Daniela Soares Ribeiro

Previsão do risco de morte de recém-nascidos prematuros de muito baixo peso

Universidade de Aveiro Departamento de Matemática
2017

Daniela Soares Ribeiro

Previsão do risco de morte de recém-nascidos prematuros de muito baixo peso

Relatório de Estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizado sob a orientação científica da Doutora Isabel Maria Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho aos meus pais Ana Soares e Manuel Ribeiro e irmão Pedro Ribeiro pelo apoio e força que me deram durante esta longa jornada, ajudando-me a alcançar um sonho de criança.
Dedico ainda ao meu avô Saul Soares que desde a minha juventude me incentivou a querer ser melhor e saber mais, dia após dia, dizendo "Estuda que eu agarro-me" .

o júri

presidente

Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais

Doutor Bruno Miguel Alves Fernandes do Gago
Professor Auxiliar Convidado do Departamento de Ciências Médicas da Universidade de Aveiro

Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

agradecimentos

Agradeço desde já à empresa MHII Solutions pela oportunidade que me proporcionou aceitando-me como estagiária e permitindo com que evoluísse o meu conhecimento numa área que me fascina, a Neonatologia. Agradeço também a todos os membros da empresa por se encontrarem sempre prontos a ajudar e por me terem feito sentir parte da família e ainda ao João Barroso e à Tânia Rocha, colegas de estágio da MHII Solutions com quem passei bons momentos de trabalho e de lazer tornando-nos numa ótima equipa.

Agradeço ainda à Professora Isabel Pereira e ao Bernardo Marques pelo conhecimento e apoio que me transmitiram juntamente com toda a sua disponibilidade.

Agradeço também ao meu padrinho de curso, Filipe Rodrigues, por me ter incentivado e ajudado em tudo, desde o primeiro momento em que o conheci. E, como não poderia deixar de ser, quero também agradecer à melhor colega de trabalho, posteriormente amiga, Filipa Oliveira, pela ótima equipa pessoal e profissional que formámos.

Por fim, mas não por último agradeço à Maggie pelas boas distrações que me proporcionou e excelente companhia durante a escrita da tese e à Paula Costa, ao Kevin Pinto e à Margarida Lopes pela força e ajuda prestada durante todo o curso. Agradeço também aos meus primos: Carolina Silva, Alexandra Santos, Diogo Silva e Rodrigo Santos pelo companheirismo, garra e aprendizagem que sempre me forneceram.

Obrigada a todos por me terem ajudado a alcançar uma grande etapa da minha vida.

Palavras Chave

Recém-nascidos, Análise de dados, Valores omissos, Regressão Logística, Seleção de Variáveis, Shiny.

Resumo

A previsão do risco de morte de recém-nascidos prematuros é um assunto de relevante importância para a tomada de decisões no âmbito da saúde pública. Com o objetivo de auxiliar os técnicos de saúde na tomada de decisões, nomeadamente no tipo de vigilância a seguir para diminuir o risco de morte dos recém-nascidos prematuros de muito baixo peso, propõe-se um modelo preditivo de regressão logística múltipla. Este modelo foi elaborado, tendo como base os dados fornecidos pela Sociedade Portuguesa de Neonatologia. O processo da construção do modelo incluiu as fases de análise e tratamento de dados, seleção de variáveis, e investigação de *outliers* e observações influentes. Para facilitar a utilização deste modelo e interpretação dos resultados correspondentes, por parte dos profissionais de saúde, foi criada uma aplicação *web*.

Keywords Newborns, Data analysis, Missing values, Logistic regression, Variables selection, Shiny.

Abstract The prediction of premature newborns is an issue of major importance in the decision making process in what public health is concerned. Aiming at helping health professionals in the decision making process, namely in the type of monitoring to follow in order to reduce the risk of death of extremely low weight premature newborns, a predictive model of multiple logistic regression is presented. This model was created according to the data made available by the Portuguese Society of Neonatology. The process of construction of the model includes analysis and data processing, the variables selection and the investigations of outliers and influent observations. To facilitate the use of this model as well as the interpretation of results, a web app was created.

Conteúdo

Lista de Figuras

Lista de Tabelas

Glossário

SPN	Sociedade Portuguesa de Neonatologia
RN	recém-nascido
ROC	características operacionais do recetor
AUC	área sob a curva
RLM	regressão logística múltipla
TIC	Tecnologias de Informação e Comunicação
AIC	critério de informação de Akaike

CAPÍTULO 1

Introdução

1.1 Introdução do Problema

De acordo com o objetivo do trabalho desenvolvido, a previsão do risco de morte de recém-nascidos prematuros de muito baixo peso, foi realizada uma análise prévia sobre modelos preditivos existentes no âmbito da predição clínica.

A predição clínica consiste em prever um determinado resultado através da análise de observações de determinadas características. As regras de predição clínica são o resultado de diversas ferramentas matemáticas utilizadas com vista a apoiar a decisão dos médicos todos os dias. Pois para eles, tomar uma determinada decisão nem sempre é fácil, visto que esta, por vezes, é demasiado complexa e acompanhada de um risco elevado.

A utilização de modelos preditivos aquando da tomada de decisão por parte dos profissionais de saúde tem imensas vantagens, visto que um modelo devidamente treinado e validado, consegue ter em conta muito mais características que o cérebro humano e não é ambíguo, sendo que, por vezes, os médicos perante um mesmo diagnóstico tomam diferentes decisões, principalmente médicos inexperientes. Porém, o facto de muitos terem receio de não conseguir utilizar os modelos preditivos faz com que estas técnicas acabem por não ser muito utilizadas, apesar da sua popularidade ter aumentado nos últimos anos.

Em 2006, Grobman e Stamilio, [1], referiram diferentes métodos para desenvolver modelos preditivos. Esses métodos poderiam ser de regressão múltipla, como a regressão linear e a regressão logística (técnicas mais tradicionais) ou utilizando outras estratégias tais como sistemas de pontos, nomogramas, árvores de decisão e redes neurais.

Os modelos de regressão múltipla englobam alguns processos comuns, tais como, a determinação das variáveis em estudo, seleção das variáveis realmente importantes, de modo a reduzir significativamente o número de variáveis (princípio da parcimônia) e análise de diferentes modelos, levando à seleção do modelo final. A regressão avalia a dependência de uma variável dependente (variável resposta, a determinar) em relação a uma variável independente (regressão simples) ou a várias (regressão múltipla). Na maioria das situações, é avaliada mais

que uma característica tendo, por isso, mais do que uma variável independente, pelo que a regressão múltipla é a mais usualmente utilizada.

A realização de testes diagnósticos para detetar a presença ou ausência de determinada doença é obtida tendo como base o modelo de regressão logística múltipla (RLM), uma vez que a variável dependente é dicotómica, ou seja, toma apenas dois valores, sendo que o 0 significa ausência e o 1 presença da característica em estudo e o modelo fornece uma probabilidade do sucesso (apresentar a doença) entre 0 e 1. Este estudo é muito utilizado, por exemplo, na pneumologia no diagnóstico de doenças de tuberculose, doenças malignas, ...

Os modelos de regressão logística múltipla são facilmente implementados existindo *softwares* próprios. No entanto, ter-se-ão de analisar se os pressupostos de aplicabilidade do modelo são satisfeitos, procedimento necessário na aplicação de qualquer modelo estatístico. Só a título de exemplo, se faltar algum valor relativo a uma variável, o modelo RLM não consegue calcular o valor correspondente da probabilidade.

Convém registar que existem outras estratégias para o desenvolvimento de modelos preditivos, tais como os sistemas de pontos, nomogramas, árvores de decisão e redes neurais.

Os sistemas de pontos consistem em atribuir pontos relativamente aos fatores, consoante sejam ou não mais significativos, através de um sistema de pontuação. A pontuação final obtida, tendo em conta todas as características avaliadas, é usada como um indicador de risco, indicando se existe um maior ou menor risco de obter um determinado resultado.

Os nomogramas são dispositivos com gráficos onde se representam relações matemáticas que ajudam o utilizador a calcular rapidamente fórmulas complicadas que demorariam algum tempo. Sendo bastante fácil de usar e em alguns casos é mais preciso do que outros modelos utilizados para o mesmo efeito. Um exemplo simples, pode ser a marcação do termómetro tendo em conta a temperatura corporal ou a previsão do pico da taxa de fluxo respiratório de pacientes asmáticos, atendendo à sua idade e altura.

As árvores de decisão têm a forma de um fluxograma sendo constituídas por perguntas em que através das respostas o paciente segue um determinado caminho, obtendo a resposta final (resultado previsto). A resposta final é a resposta ao teste em questão podendo ser apresentada sobre a forma qualitativa (sim ou não, presença ou ausência, respetivamente) ou forma quantitativa (sob a forma de percentagem). As suas vantagens são a apresentação gráfica de fácil interpretação e compreensão e, ainda o facto da sua construção ser possível sem depender do conhecimento dos dados de todos os pacientes.

Existem ainda as redes neurais, sendo algo já mais sofisticado, onde é possível usar um grande número de variáveis e o modelo aprender com a estrutura dos dados. Pois ele recebe todas essas variáveis (*inputs*), processa toda essa informação e, de seguida, tendo em conta essas características fornece um resultado final (*output*). Estas redes têm uma enorme vantagem, uma vez que conseguem aprender a estrutura dos dados tendo em conta apenas as variáveis de entrada e a resposta correspondente, isto é conseguido fornecendo ao programa uma base de dados completa com os dados de entrada e o resultado final associado. Visto que o computador trata basicamente de todo o processo são necessários bons recursos computacionais. Sendo uma técnica recente, ainda não está suficientemente explorada na

literatura.

Concluindo, existem diversos métodos de previsão, cada um com as suas vantagens e desvantagens correspondentes. Cabe apenas ao investigador selecionar o mais adequado a cada situação, tendo em conta todos os objetivos em questão, a validade dos pressupostos do modelo estatístico, o grau de complexidade do método e a simplicidade de interpretação dos resultados. Um aspeto a ter em linha de conta na escolha da metodologia a seguir, além dos aspetos anteriormente referidos, é que ela incentive a sua utilização por parte dos técnicos de saúde, diminuindo o tempo da tomada de decisão e consequentemente aumentando o tempo de reação. (Adams e Leveson, [2]; Neves e Dias, [3])

Para se ter alguma sensibilidade e fazer uma adequada escolha da metodologia mais apropriada tendo em linha de conta os aspetos anteriormente referidos, foram analisados alguns artigos que abordavam o problema em causa, a morte de recém-nascidos prematuros de muito baixo peso.

Os artigos analisados recorriam, na sua grande maioria, ao modelo de regressão logística múltipla para a previsão do risco de morte, com recurso ao método de *stepwise* para a seleção de variáveis, ao cálculo do valor sob a curva ROC (*Receiver Operating Characteristic*) para avaliar a qualidade das previsões e ainda à utilização do teste de Hosmer & Lemeshow para evidenciar (ou não), a adequação do modelo encontrado (Marshall *et al*, [4]). E estes dois últimos eram utilizados também com o objetivo de fazer a comparação de modelos (Pollack *et al*, [5]) .

Na maioria dos artigos, os dados eram referentes a várias Unidades de Cuidados Intensivos Neonatais e relativos a um longo período de tempo (mínimo de 1 ano), nos quais o peso do recém-nascido era inferior a 1500 ou 2500 gramas e a idade gestacional inferior a 32 ou 37 semanas.

Tendo em conta a sua influência na morte, as variáveis analisadas foram variáveis maternas como, por exemplo, a idade da mãe, a nacionalidade, o grau de escolaridade, o uso de esteroides e o número de consultas pré-natais realizadas, a gestação múltipla, o local do parto e o tipo de parto. E ainda, variáveis neonatais como, por exemplo, o sexo do bebé, a idade gestacional, o peso, o percentil 10 (se se encontrava acima ou abaixo desse percentil, (Lim *et al*, [6])), o Apgar1, o Apgar5, a presença de anomalia congénita e ressuscitações realizadas. Foi também analisada a existência de glóbulos vermelhos nucleados (verificando um aumento entre o segundo e quinto dia, (Cremer *et al*, [7])). Sendo que o peso era uma das principais características da causa de morte.

Em relação às variáveis analisadas anteriormente, concluiu-se que a idade da mãe (idades de adolescência e menopausa), os valores baixos de Apgar - indicador do estado do recém-nascido aquando do nascimento, o reduzido número de consultas pré-natais, o parto vaginal, a presença de anomalia congénita, o sexo masculino, reduzido peso e idade gestacional afetavam negativamente o estado do recém-nascido (RN), levando ao aumento da probabilidade de morte (Marshall *et al*, [4]; Aparecida *et al*, [8]) .

A prematuridade é algo que se poderia evitar prestando mais e melhores cuidados de saúde à mãe e ao bebé, acompanhando o peso materno, eliminando o tabagismo e tendo em conta a baixa escolaridade da mãe, pois a existência de baixos recursos económicos poderá levar a uma má alimentação da mãe e, consequentemente, à perda acentuada do peso do RN (Aparecida *et al*, [8]).

Cunha *et al*, [9], fizeram um estudo do estado dos recém-nascidos de baixo peso (igual ou inferior a 1000 gramas) aos dois e três anos de idade, tendo em conta o seu crescimento e desenvolvimento, pois as crianças prematuras de baixo peso apresentam um elevado risco de sequelas. Este estudo foi realizado com base em resultados do registo nacional de muito baixo peso de 2005 e 2006. Sendo que o objetivo era encontrar variáveis comuns a todas as unidades que ao serem aplicadas pudessem diminuir e evitar os efeitos do baixo peso e prematuridade, percebendo ainda as necessidades especiais que estas crianças necessitam nos seus primeiros anos de vida, diminuindo o seu risco de morte e melhorando a sua qualidade de vida. Foi também avaliada a situação familiar (o RN vivia sobretudo com a família e esta tinha um baixo grau de escolaridade), além da existência de várias sequelas associadas, tais como surdez, cegueira, problemas neurológicos, paralisia cerebral e atraso de desenvolvimento. Recorreu-se à regressão logística para prever o risco de paralisia cerebral ou atraso de desenvolvimento. Sendo que os recém-nascidos que apresentavam melhores resultados eram os que continuavam a ser acompanhados, após a alta, pela Unidade de Cuidados Intensivos Neonatais.

Portugal, devido aos investimentos que tem na saúde neonatal, é um dos países com taxas mais baixas de mortalidade infantil e neonatal da Europa. De acordo ainda com estes autores (Cunha *et al*, [9]), é importante não só o conhecimento de como evitar a prematuridade, mas também perceber os seus efeitos, como o crescimento e desenvolvimento de sequelas, permitindo assim encontrar uma solução, de modo a encontrar estratégias de diagnósticos de sequelas e medidas de atuação.

1.2 Estrutura do Relatório

O relatório é constituído por cinco capítulos que seguem uma ordem lógica de modo a explicar todo o trabalho desenvolvido durante este último ano curricular, tanto de estágio como de pesquisa.

No Capítulo 1 é apresentada a introdução ao problema acompanhada por toda a revisão bibliográfica e, de seguida, toda a estrutura do relatório.

No Capítulo 2 descreve-se a empresa responsável pelo tema do relatório e indicam-se os objetivos do trabalho desenvolvido. Visto que o estudo efetuado é centrado nos recém-nascidos prematuros é ainda fornecida uma breve introdução sobre a Neonatologia.

Para a previsão do risco de morte foi necessário o estudo e desenvolvimento de um modelo de regressão logística, pelo que é apresentada no Capítulo 3 a teoria associada, juntamente com a validação do modelo e todos os processos envolvidos. Tendo-se recorrido maioritariamente às seguintes referências bibliográficas (Marôco, [10]; Hall *et al*, [11]; Turkman e Silva, [12]; DeMaris, [13]).

Todos estes capítulos serviram de apoio para o objetivo final, que consistiu em tratar a base de dados fornecida pela Sociedade Portuguesa de Neonatologia (SPN) de modo a sustentar um modelo com base neles que calculasse o risco de morte em tempo real. No Capítulo 4 é então apresentado todo o processo de análise de dados, desde a análise da base de dados, passando pelo tratamento de valores não recodificados e omissos, à seleção de variáveis a utilizar, à obtenção do modelo final e ainda tratamento de *outliers* e observações influentes. No final é ainda obtida uma aplicação *web* constituída por um formulário em que ao preencher fornecerá o tal risco de morte tão aguardado.

O relatório termina com o capítulo final, Capítulo 5 onde são apresentadas todas as conclusões obtidas com a realização deste trabalho, algumas críticas e ainda possíveis propostas de um trabalho futuro.

CAPÍTULO 2

Empresa versus Neonatologia

2.1 Introdução

O relatório foi desenvolvido no âmbito de um estágio curricular do mestrado em Matemática e Aplicações na área de especialização em Estatística e Investigação Operacional, efetuado na empresa MHII Solutions. Este estágio tinha o objetivo de aplicar modelos preditivos no contexto da Neonatologia, tendo em conta os dados fornecidos pela SPN.

A MHII Solutions tem parcerias com várias instituições em Portugal e no resto do mundo. De uma forma particular a parceria estabelecida com a SPN tem como principal objetivo analisar a morte dos bebés prematuros de muito baixo peso, de modo a que os profissionais de saúde possam estar mais vigilantes e assim, diminuírem o risco de morte, não só através dos seus cuidados mas também das previsões realizadas pela empresa.

2.2 Descrição da Empresa

A MHII Solutions foi fundada em Setembro de 2015 e apesar de estar localizada em São João da Madeira já tem uma presença internacional. É o resultado do foco e experiência adquirida num dos mais importantes setores: a saúde. Sendo constituída por uma equipa multidisciplinar com experiência em *data engineering*, *advanced analytics*, *business intelligence* e *interaction design* que tem por objetivo fornecer soluções que melhorem a saúde e o bem-estar da população.

Preocupa-se em resolver problemas técnicos e desafiantes relacionados com a saúde, por exemplo, diminuição da fila de espera nas urgências e do risco de morte de bebés prematuros, fornecendo soluções e melhorando, assim, os processos de tomada de decisão. Trabalha com instituições públicas e privadas de saúde como, por exemplo, hospitais, clínicas, ONG's, ...

Os seus maiores desafios consistem em assegurar o retorno do investimento das organizações de saúde, trazendo valor acrescentado aos cuidados de saúde e apoiando ativamente as diferentes partes interessadas.

Tem como missão a melhoria dos cuidados de saúde, aumentando a qualidade de vida e o bem-estar das pessoas, diminuindo o risco de morte e aumentando a esperança média de vida. Por outro lado, pretende também auxiliar os profissionais de saúde, uma vez que tem em vista um conjunto de procedimentos testados para melhor fazer a monitorização e vigilância do estado de saúde do doente.

Como dito anteriormente, já tem uma presença internacional pois não se encontra apenas em Portugal (sede) mas no resto do mundo como França, Quênia, Estados Unidos e Brasil, onde está a criar uma plataforma analítica com o objetivo de tratar indicadores de recém-nascidos prematuros, monitorizar dados nacionais sobre a HIV/SIDA, visualizar dados de mutilação genital feminina e antecipar a disseminação mundial do vírus Zika, respetivamente.

Recentemente a MHII Solutions fundiu-se com a Prologica, uma das mais antigas empresas portuguesas que operam na área das Tecnologias de Informação e Comunicação (TIC) desde 1984.

2.3 Introdução à Neonatologia

Neonatologia deriva do latim (ne(o) - novo , nat(o) - nascimento e logia - estudo) tendo por base o conhecimento do recém-nascido. É um ramo da Pediatria onde se prestam cuidados aos recém-nascidos desde o seu nascimento até aos 28 dias. Apesar de ser uma especialidade relativamente recente (século XIX), tem vindo a proporcionar ótimos resultados na saúde neonatal pois com o conhecimento e compreensão da fisiologia dos bebés prematuros e do funcionamento dos seus órgãos e criação de unidades de cuidados intensivos, tem vindo a diminuir a taxa de mortalidade neonatal. (XXS, [14])

Duas das características de um recém-nascido que influenciam o seu normal desenvolvimento são a idade gestacional e o peso, encontrando-se relacionadas. A idade gestacional é o número de dias decorridos entre o primeiro dia do último período menstrual normal e o dia do nascimento, tendo como duração média 280 dias ou 40 semanas. O peso, que em média tem o valor de 3200 gramas, deve ser medido logo na sala de partos, pois a perda de peso ocorre rapidamente.

Existem três tipos de classificações do RN relativamente à idade gestacional: pré-termo (inferior a 37 semanas), de termo (entre 37 a 42 semanas) e pós-termo (superior a 42 semanas).

Considerando apenas o RN de termo e analisando o seu peso, este pode ser classificado como (Araujo e Reis, [15]):

• pequeno ou leve para a idade gestacional (peso inferior ou igual a 2500 gramas encontrando-se abaixo do percentil 10 com atraso de crescimento intra-uterino);

• peso adequado para a idade gestacional (entre 2500 e 3800/4000 gramas encontrando-se entre o percentil 10 e 90, apresentando um crescimento normal);

• grande para a idade gestacional (peso superior ou igual a 3800/4000 gramas, estando acima do percentil 90, crescendo rapidamente e podendo ter desenvolvido esse padrão acelerado durante a gravidez).

Algumas características do RN são não só a idade gestacional e o peso, como referido anteriormente, mas também o comprimento e perímetro cefálico, podendo ser medidos mais tarde, sendo para o segundo o mais correto no dia seguinte. Para o RN de termo o comprimento encontra-se entre 48 a 53 cm e o perímetro cefálico entre 31 e 35 cm.

Evidentemente que nenhuma gravidez se encontra livre de problemas ou contratempos. É de evidenciar que o recurso mais generalizado a tratamentos de fertilidade e a idade mais avançada das grávidas aquando do primeiro filho, aumenta o risco de uma gravidez prematura (Santos, [16]). Este é um dos problemas da atualidade, pelo que se torna mais necessário fornecer cuidados aos RN prematuros de modo a melhorar a qualidade de vida deles, diminuindo o seu risco de morte.

Segundo o Instituto Nacional de Estatística relativamente a um estudo realizado sobre as Estatísticas Demográficas de 2015, entre 2010 e 2015, o número de RN de baixo peso (peso inferior a 2500 gramas) e prematuros (idade gestacional inferior a 37 semanas) tem vindo a aumentar. Em 2015, os RN de baixo peso representavam uma média de 8.9% e os RN prematuros uma média de 8% dos nascimentos vivos com valores acima destes para idades das mães inferiores a 20 anos e superiores a 34 anos (INE p.51-52, [17]).

Os recém-nascidos são designados prematuros ou de pré-termo (classificação referenciada anteriormente) quando têm uma idade gestacional inferior a 37 semanas, pois normalmente uma gravidez dura entre 37 a 42 semanas (recém-nascido de termo).

Tal como visto anteriormente, duas das características de um bebé prematuro são o baixo peso e o tamanho pequeno, ou seja, relativamente aos percentis, estes valores encontram-se muitas vezes abaixo do percentil 10.

Para uma idade gestacional inferior a 37 semanas e tendo em conta também o peso, um bebé prematuro pode ser classificado em (XXS, [14]):

• Pré-Termo Limiar: com idade gestacional entre 33 e 36 semanas e/ou peso à nascença entre 1500g e 2500g.

• Prematuro Moderado: com idade gestacional entre 28 e 32 semanas e/ou peso à nascença entre 1000g e 2500g.

• Prematuro Extremo: com idade gestacional inferior a 28 semanas e/ou peso inferior a 1000g.

O prematuro extremo é aquele que apresenta problemas mais frequentes e mais graves por esse mesmo motivo (elevada imaturidade).

De modo a reduzir as complicações neonatais associadas ao nascimento prematuro é recomendada a ingestão de corticoides pré-natais, pois estes reduzem significativamente tanto o risco de morte como da ocorrência de síndrome de dificuldade respiratória e de hemorragia intra-ventricular. Quando recomendado, geralmente é administrado um único ciclo de corticoides, uma vez que apenas um ciclo não cria complicações posteriores, nem

maternas nem fetais e a sua administração pode variar entre a 23ª e a 39ª semana de gestação dependendo de cada caso em específico (SPP, [18]).

A primeira avaliação do estado do recém-nascido é realizada imediatamente após o parto, 1º, 5º e 10º minutos, de modo a detetar a existência de casos estranhos que necessitem de assistência imediata (1º minuto) e avaliar os resultados (5º minuto). Esta avaliação tem o nome de Apgar1, Apgar5 e Apgar10, respetivamente. Para a determinação do índice de Apgar são avaliados cinco parâmetros: a frequência cardíaca, a respiração, o tónus muscular, a resposta a estímulos e a cor da pele. Pelo que, a frequência cardíaca é referente ao movimento cardíaco indicando ausência caso não haja, < 100 se for inferior a 100 e >100 caso seja superior a 100. A respiração pode ser ausente, caso não haja, fraca se o choro for fraco e boa caso o choro seja vigoroso. O tónus muscular refere-se à capacidade de movimentação que o bebé apresenta. A resposta a estímulos avalia a capacidade de reação do bebé e a cor da pele indica a presença (rosado) ou ausência (cianose, ou seja, cor azulada) de oxigênio no sangue do bebé. Cada parâmetro tem uma pontuação máxima de 2 pontos e mínima de 0 de acordo com a qualidade do seu estado. Sendo assim, o índice de Apgar pode tomar valores entre 0 e 10. Valores de 0 a 3 significa que o RN se encontra com dificuldade severa, de 4 a 6 com dificuldade moderada e de 7 a 10 sem dificuldade. A tabela 2.1 sintetiza a informação anteriormente referida em termos dos aspetos avaliados na determinação do índice de Apgar (Araujo e Reis, [15]).

Concluindo, existem diversas formas de avaliar um RN mas tão importante ou mais que esta avaliação é a capacidade de decisão e reação perante algo fora do normal e isso engloba todos os processos com que os profissionais de Neonatologia se deparam todos os dias. Esta especialidade e trabalho são de extrema importância no dia-a-dia de todas as famílias, pois muitas das vezes o futuro do RN depende deles.

Parâmetros	0	1	2
Frequência Cardíaca	Ausente	< 100	> 100
Respiração	Ausente	Fraca, irregular	Choro vigoroso
Tónus muscular	Hipotonia	Ligeiro	Ativo, com tonicidade
Resposta a estímulos	Ausente	Caretas	Choro vigoroso
Cor da pele	Pálido/cianose	Rosado, extremidades cianosadas	Todo rosado

Tabela 2.1: Aspetos avaliados na determinação do índice de Apgar

CAPÍTULO 3

Modelo de Regressão Logística

3.1 Introdução

A realização do relatório tem como objetivo a obtenção de um modelo de regressão logística múltipla para a previsão do risco de morte de recém-nascidos prematuros de muito baixo peso à nascença. Um modelo de regressão logística é um modelo linear generalizado, isto é, um modelo em que a variável dependente (variável resposta) é determinada à custa das variáveis independentes (variáveis fornecidas pela base de dados). A variável dependente toma apenas valores discretos e é dicotómica (tendo apenas dois valores), sendo por isso um modelo discreto. A regressão logística tem como objetivo avaliar a dependência da variável dependente relativamente a todas as variáveis independentes, determinando portanto, no caso em estudo, o risco de morte a partir das variáveis iniciais, obtidas até aos primeiros dez minutos de vida, que caracterizam o recém-nascido.

3.2 Classificação de Variáveis

Um conjunto de indivíduos com as mesmas características em estudo designa-se de população, onde cada característica é denominada de atributo e apresentada sob a forma de variável. Todas estas características descrevem de certa forma uma determinada observação.

As variáveis podem ser classificadas de duas formas distintas consoante a sua natureza, podendo ser qualitativas (categóricas) ou quantitativas (McCall, [19]).

Apesar de poderem ser codificadas por números, as variáveis qualitativas representam categorias, sendo variáveis de texto, fatores. Por exemplo, a variável Tipo de Parto é definida pelas categorias Vaginal ou Cesariana, podendo-se atribuir o número 1 para representar o parto Vaginal e o 2 para o parto por Cesariana. Porém, este tipo de variáveis também podem apresentar uma certa ordem, por exemplo, na variável Corticoides Pré-natais (1 significa Não, 2 Parcial e 3 Completo). Olhando para estes dois tipos de variáveis conclui-se que existem dois tipos de escalas para o atributo, no primeiro é a escala nominal e no segundo é a escala ordinal.

Já as variáveis quantitativas são aquelas em que a sua natureza é numérica pelo que cada número vale por si só e não tem correspondência (legenda ou recodificação). Por exemplo, o Perímetro Cefálico ou o Apgar. Porém, a primeira toma valores contínuos, ou seja, pode tomar qualquer valor dentro de um intervalo de números reais, enquanto que a segunda toma valores discretos, em número finito ou quando for infinito será numerável. Neste caso o Perímetro Cefálico assumia valores no intervalo [2.8, 40.0] e o Apgar tomava valores inteiros entre 0 e 10.

3.3 Modelo Linear Generalizado

Suponhamos que estamos interessados em estudar uma característica representada pela variável aleatória Y, que designamos por variável resposta ou dependente, a qual depende de p variáveis explicativas, também designadas por covariáveis ou variáveis independentes. Admitimos que temos n observações na amostra. Assim, a representação matricial do modelo linear generalizado é definida por

$$Y = X\beta + \varepsilon,$$

onde

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

e os erros aleatórios ε_i são não correlacionados, de valor médio nulo e variância constante. Tipicamente assume-se que $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$.

Existem vários tipos de Modelos Lineares Generalizados, de entre os quais o modelo de regressão linear, de análise de variância e covariância, de regressão logística, de regressão de Poisson, etc. Considerando os modelos de regressão linear e regressão logística, a maior diferença entre eles é a natureza da variável dependente (variável resposta); no primeiro a variável é contínua (podendo tomar qualquer valor), já no segundo é discreta, sendo na maioria das vezes dicotómica tomando os valores 0 e 1 (caso em estudo), porém pode também ser policotómica (tomando assim mais do que dois valores).

3.4 Regressão Linear versus Regressão Logística

O valor médio de Y dado X, E(Y|X) é muito importante na análise de regressão, visto que o grande objetivo da regressão é avaliar a dependência da variável dependente relativamente a todas as variáveis independentes, determinando assim o valor da variável dependente. Na regressão linear, E(Y|X) pode assumir qualquer valor, já na regressão logística apenas pode tomar os valores 0 ou 1.

O modelo de regressão linear simples é do tipo $y_i = \beta_0 + \beta_1 x_i$, $i = 1, \dots, n$. Na regressão linear simples o valor médio de Y_i, condicional a x_i é dado por:

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i.$$

Logo, tomando apenas por uma questão de simplicidade $\pi(x) = E(Y|x)$ tem-se $y = \pi(x) + \varepsilon$. Uma vez que na regressão logística, os valores possíveis para y são 0 ou 1, tem-se:

$$\begin{cases} 0 = \pi(x) + \varepsilon \\ 1 = \pi(x) + \varepsilon \end{cases} \Rightarrow \begin{cases} \varepsilon = -\pi(x) \\ \varepsilon = 1 - \pi(x) \end{cases}.$$

Isto é,

$$\begin{cases} \varepsilon = -\pi(x) \text{ com probabilidade } 1 - \pi(x), y = 0 \\ \varepsilon = 1 - \pi(x) \text{ com probabilidade } \pi(x), y = 1 \end{cases}.$$

Determinação do valor da média e variância do erro associado:

$$E(\varepsilon) = (1 - \pi(x))\pi(x) + (-\pi(x))(1 - \pi(x)) = 0,$$

$$V(\varepsilon) = E(\varepsilon^2) = (1 - \pi(x))^2\pi(x) + (-\pi(x))^2(1 - \pi(x)) = (1 - \pi(x))\pi(x).$$

Logo, $\varepsilon \sim \mathcal{N}(0, (1 - \pi(x))\pi(x))$.

3.5 Regressão Logística Simples

Um modelo de regressão logística simples é um modelo linear generalizado em que a variável dependente é determinada apenas à custa de uma variável independente. A forma de $\pi(x)$ é traduzida por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Linearizando $\pi(x)$, ou seja, fazendo a transformação *logit* vem:

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \ln(e^{\beta_0+\beta_1 x}) = \beta_0 + \beta_1 x.$$

Esta transformação é importante pois o modelo passa a possuir diversas propriedades do modelo de regressão linear.

3.5.1 Ajustamento do Modelo

Para se fazer o ajustamento do modelo é necessário proceder à estimação dos parâmetros (β_0, β_1). O método usado é o Método de Máxima Verosimilhança.

Tendo em conta que $P(Y = 1) = \pi(x)$ e $P(Y = 0) = 1 - \pi(x)$, isto é, $Y \sim \mathcal{B}(1, \pi(x))$, a função massa de probabilidade para Y calculada em $Y = y_i$, é dada por:

$$f(y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i)^{1-y_i}, y_i \in \{0, 1\}.$$

Assumindo independência entre observações a verosimilhança é definida por

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, y_i = 0, 1, i = 1, \ldots, n.$$

Logaritmizando vem

$$L(\beta) = ln(l(\beta)) = ln\left[\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}\right] = \sum_{i=1}^{n}[y_i ln(\pi(x_i)) + (1-y_i)ln(1-\pi(x_i)))].$$

Uma vez que

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},$$

tem-se que

$$L(\beta) = \sum_{i=1}^{n}\left[y_i ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) + (1-y_i)ln\left(1-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right] = \sum_{i=1}^{n} y_i\left[(\beta_0+\beta_1 x_i) - ln(1+e^{\beta_0+\beta_1 x_i})\right].$$

Para obter as estimativas de máxima verosimilhança para os parâmetros ter-se-á de encontrar os valores de β_0 e β_1 que maximizem $L(\beta)$.

3.6 Regressão Logística Múltipla

A única diferença entre a regressão logística simples e múltipla é o número de variáveis independentes, pelo que a última tem no mínimo duas. Sendo assim, todos os pressupostos referidos anteriormente manter-se-ão porém agora com $x = (x_1, x_2, \ldots, x_p)$.

Tem-se então

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}.$$

Linearizando $\pi(x)$, ou seja, fazendo a transformação *logit* vem:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

3.6.1 Ajustamento do Modelo

O ajustamento do modelo consiste na estimação dos parâmetros do modelo $(\beta_0, \beta_1, \cdots, \beta_p)$. O método mais usual é o Método da Máxima Verosimilhança, que como se referiu anteriormente, se baseia na verosimilhança da amostra. A log-verosimilhança é dada por

$$L(\beta) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - ln(1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}).$$

Para se obterem os estimadores de máxima verosimilhança ter-se-á de resolver um sistema de $p+1$ equações constituídas pelas derivadas parciais da log-verosimilhança relativamente a cada um dos parâmetros.

3.7 Significância e Qualidade do Modelo

A avaliação da significância e qualidade do modelo é realizada com base em três testes distintos: Teste de Wald, Análise de Variância e Teste de Hosmer & Lemeshow, sendo que os dois primeiros avaliam a significância dos coeficientes do modelo e o último avalia a sua qualidade.

3.7.1 Teste de Wald

O objetivo do teste de Wald é a determinação das variáveis independentes que influenciam significativamente a variável dependente. Hosmer e Lemeshow, [20], definem algumas etapas importantes a considerar na construção de um modelo de regressão logístico múltiplo. Começa-se por identificar as variáveis independentes candidatas a entrar no modelo, fazendo uma avaliação individual e considerando o modelo de regressão logístico simples. Tendo identificado as variáveis independentes constrói-se o modelo de regressão logístico múltiplo, avaliando-se a sua importância através do teste de Wald. Pretende testar a significância dos coeficientes do modelo, ou seja, determinar se considerando todos os outros coeficientes exceto um determinado (a testar) ele é ou não significativo, isto é, se não pode ou pode ser considerado nulo, respetivamente. Tendo isto em conta as hipóteses estatísticas são:

$$H_0 : \beta_i = 0 | \beta_0, \beta_1, \beta_{i-1}, \beta_{i+1}, \beta_p$$

$$vs$$

$$H_1 : \beta_i \neq 0 | \beta_0, \beta_1, \beta_{i-1}, \beta_{i+1}, \beta_p, i = (1, ..., p).$$

A estatística de teste é dada por:

$$T_{wald_i} = \frac{\hat{\beta}_i}{\hat{SE}\left(\hat{\beta}_i\right)} \overset{a}{\sim} \mathcal{N}(0,1),$$

onde $\hat{\beta}_i$ é o estimador de β_i (parâmetro) e $\hat{SE}(\hat{\beta}_i)$ é o estimador do erro padrão de β_i. Esta estatística tem uma distribuição assintoticamente Normal. H_0 é rejeitado, ou seja, β_i é considerado significativo quando $p\text{-}value \leq \alpha$, sendo α um valor pequeno.

O modelo múltiplo construído deve sempre ser comparado com o anterior, aplicando-se o teste da razão de verosimilhanças. Este processo de retirar e reajustar o modelo é repetido até se verificar que as variáveis explicativas incluídas no modelo descrevam adequadamente o comportamento da variável resposta.

Concluindo, este teste é considerado mais fiável para dimensões de amostras elevadas (a estatística de teste segue uma distribuição assintótica) ou para coeficientes pequenos, pois caso contrário pode levar à inflação dos resultados do SE levando à não rejeição de H_0. Caso isto se verifique deve-se fazer um teste de rácio de verosimilhanças.

3.7.2 Teste de Análise de Variância

A análise de variância, mais conhecida por ANOVA, é utilizada com o objetivo de comparar a influência de vários fatores (variáveis independentes) na resposta em questão (variável dependente), obtendo assim os mais influentes. Se o número de fatores a ter em conta for um, a análise da variância é designada de ANOVA *one-way*, caso seja superior a um, será denominada de ANOVA fatorial.

A ANOVA é constituída por duas técnicas: a análise de variância e o teste de Kruskal-Wallis. Sendo que a primeira é utilizada quando os grupos são modelados por distribuições

normais de igual variância, caso isto não aconteça, é utilizada a segunda técnica, técnica não paramétrica. A única diferença entre as duas é a medida de localização.

Cada observação da variável independente é designada de nível de fator e a combinação dos níveis de todos os fatores é denominada de tratamento. A ANOVA diz-se de efeitos fixos quando estes níveis são fixados à partida, se forem selecionados aleatoriamente, a ANOVA diz-se de efeitos aleatórios. Sendo que, a ANOVA a utilizar no desenvolvimento deste trabalho será a ANOVA fatorial, ou seja, uma ANOVA com vários níveis pelo que existem imensas variáveis independentes. Inicialmente, será abordada a ANOVA a um fator e a dois, de modo a conseguir transpor a ideia para o caso em que o número de fatores seja superior a dois, caso em estudo.

ANOVA com um Fator e Efeitos Fixos

Neste caso, tal como referido anteriormente, o comportamento da variável dependente (y) é influenciado por apenas um fator, constituído por k grupos (k observações diferentes). A variável y pode ser explicada pelo seguinte modelo (Barnes, [21]):

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

em que y_{ij} representa a observação i da amostra j com (i=1, ..., n_j; j=1, ..., k) e n_j é a dimensão da amostra j. Sendo por isso, a dimensão total da amostra $N = \sum_{i=1}^{k} n_i$. Tem-se ainda que μ é a medida global, α_i o efeito do tratamento para a amostra i e ε_{ij} um erro aleatório associado a cada observação ij.

A ANOVA tem como pressupostos que a distribuição de y seja normal, as variâncias populacionais homogêneas, resultando o facto dos erros ε_{ij} serem independentes e seguirem uma distribuição normal com média 0 e variância σ^2, ou seja, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

As hipóteses a serem testadas são, o facto das médias populacionais diferirem ou não significativamente entre si:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$vs$$

$$H_1 : \exists i, j : \mu_i \neq \mu_j, (i \neq j; i, j = 1, ..., k, k \geq 2)$$

ou, equivalente, o fator em estudo (variável independente) ter ou não significância para a variável dependente:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$$

$$vs$$

$$H_1 : \exists i : \alpha_i \neq \mu_j, (i = 1, ..., k).$$

Se os efeitos (níveis do fator) forem significativamente nulos, significa que as médias são significativamente semelhantes.

Tendo em conta o modelo teórico da ANOVA na população é possível obter o seguinte modelo a partir das observações amostrais:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i),$$

onde $\bar{y}$ é a média geral da amostra, $(\bar{y}_i - \bar{y})$ é o efeito do tratamento e $(y_{ij} - \bar{y}_i)$ os resíduos.

Tem-se que

$$SQE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 ,$$

$$SQF = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

e

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 ,$$

onde SQE é a soma dos quadrados dos erros, SQF a soma dos quadrados do fator e SQT a soma dos quadrados totais, estando relacionados através de SQT = SQF + SQE.

A estatística de teste é dada por:

$$F = \frac{MQF}{MQE} \frown F_{k-1,N-k},$$

sendo que F tem a distribuição de Fisher, onde sob hipótese de H_0 tem-se que $\frac{SQF}{\sigma^2} \sim \chi^2_{k-1}$ e que $\frac{SQE}{\sigma^2} \sim \chi^2_{N-k}$.

A hipótese nula será rejeitada para valores demasiado elevados da estatística de teste, pois quando as médias não são todas iguais há tendência para inflacionar o numerador.

ANOVA a dois e mais Fatores

Consideremos o caso de dois fatores. O modelo é dado por:

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr},$$

onde o fator A tem a níveis (i=1, ..., a) e o fator B tem b níveis (j=1, ..., b). Cada combinação dos níveis dos fatores A e B possui r repetições (supondo que as amostras têm a mesma dimensão), α_i é o efeito do fator A, β_j o efeito do fator B, γ_{ij} representa a interação entre os dois fatores, μ a média global e ε_{ijr} os erros, em que $\varepsilon_{ijr} \sim \mathcal{N}(0, \sigma^2)$.

Tendo em conta o modelo teórico da ANOVA na população tem-se o modelo a partir das observações amostrais:

$$y_{ijr} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}) + (y_{ijr} - \bar{y}_{ij}) .$$

Uma vez que o objetivo é testar se para cada nível do fator A e B as médias são ou não iguais (se forem iguais estes fatores não têm um efeito significativo), têm-se as seguintes hipóteses:

$$H_0^A : \mu_1 = \mu_2 = ... = \mu_a$$

$$vs$$

$$H_1^A : \exists i, j : \mu_i \neq \mu_j, (i \neq j; i, j = 1, ..., a);$$

$$H_0^B : \mu_1 = \mu_2 = ... = \mu_b$$

$$vs$$

$$H_1^B : \exists i, j : \mu_i \neq \mu_j, (i \neq j; i, j = 1, ..., b);$$

e

$$H_0^\gamma : \gamma_{ij} = 0$$

$$vs$$

$$H_1^\gamma : \gamma_{ij} \neq 0 (i = 1, ..., a; j = 1, ..., b)$$

para testar se existe interação entre os fatores.

Tem-se

$$SQT = SQF_A + SQF_B + SQ_{AxB} + SQE,$$

onde SQF_A é a soma dos quadrados do fator A, SQF_B é a soma dos quadrados do fator B e SQ_{AxB} é a soma dos quadrados da interação. Atendendo a que se terão de considerar três estatísticas de teste, o que já começa a ser demasiado trabalhoso, torna-se imprescindível recorrer a um *software* estatístico. Facilmente se teria a ANOVA com três ou mais fatores através da extensão da ANOVA de dois fatores.

3.7.3 Teste de Hosmer & Lemeshow

O teste de Hosmer & Lemeshow é um teste de ajustamento do modelo, ou seja, é um teste realizado com o objetivo de ver se o modelo se ajusta ou não aos dados (Hosmer e Lemeshow, [20]).

Esta estatística de teste tem por base a realização de um teste de χ^2 aplicada a uma tabela de contingência $2 \times g$. Esta tabela é construída classificando as duas classes da variável dependente dicotómica por g grupos definidos pelos decis das probabilidades de sucesso.

A estatística de teste é dada por:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i},$$

onde g é o número de grupos, O_i o número de sucessos observados da variável dependente em cada classe dos grupos e E_i o valor esperado nessas classes.

A estatística χ^2_{HL} para amostras de elevada dimensão tem distribuição assintótica χ^2 com $g-2$ graus de liberdade. Rejeita-se a hipótese do modelo se ajustar aos dados para um *p-value* $\leq \alpha$, sendo α um valor de significância pré-determinado e significativamente pequeno.

As probabilidades de sucesso são, geralmente, distribuídas por dez grupos, onde a probabilidade da variável resposta de cada observação é distribuída por esses dez grupos do seguinte modo: [0, 0.1[, [0.1, 0.2[, ..., [0.9, 1]. É dito que um modelo se ajusta bem aos dados quando os sucessos se encontram nas classes referentes a maiores probabilidades e os insucessos nas classes mais baixas, de probabilidades inferiores. Ora, o facto do modelo se ajustar aos dados

significa que os valores das probabilidades observadas são relativamente próximos dos valores das probabilidades esperadas. Caso o ajustamento do modelo não seja tão bom significa que os sucessos e insucessos se encontram dispersos pelos grupos de classes.

3.8 Tratamento dos Valores Omissos

A existência de valores omissos aquando da análise de dados é um dos maiores e desafiantes problemas, visto que é totalmente diferente ter dados reais ou obtidos através de qualquer processo que se possa utilizar. Pois, por muito bom que seja o método encontrado ele nunca será preferível a uma base de dados completa tida inicialmente (Vach, [22]).

No processo de tratamento de valores omissos o mais importante, inicialmente, nem é tanto o tratamento mas sim o modo como evitar o seu aparecimento, visto que a qualidade dos resultados é influenciada pela recolha e características dos dados. Apesar destes serem muitas vezes constituídos por variáveis qualitativas e quantitativas, é possível imputar apenas os valores omissos das variáveis quantitativas, que são muitas vezes, estes os valores omissos com os quais os programas não conseguem lidar.

Quando existe uma grande quantidade de dados é frequente a existência de valores omissos. Porém, há sempre formas de os evitar. Por exemplo, já existem programas informáticos para o preenchimento de dados que não permitem deixar nada em branco, ou seja, não deixam passar para a variável seguinte se a anterior não tiver sido preenchida.

A existência de valores omissos pode dever-se a vários fatores como, por exemplo, o tipo de estudo, o processo de amostragem e o objetivo do inquérito. Porém, no presente estudo deve-se à falta de preenchimento aquando do registo, provavelmente por esquecimento ou falta de informação. O maior problema da existência de valores omissos não é a falta de informação e todo o transtorno que isso causa mas o condicionamento da análise estatística, pois uma má escolha do método poderá colocar em causa toda a análise realizada. (Pereira, [23])

Antes de se optar por qualquer método, deverá ter-se em conta as variáveis a utilizar, a percentagem de valores omissos na totalidade e em cada variável. Se existirem variáveis com demasiados valores omissos que não sejam explicativas em relação à variável dependente não valerá a pena considerá-las. Caso a percentagem de valores omissos na totalidade seja muito reduzida e esses valores não dependerem de outros, isto é, sejam omissos aleatoriamente poderá aplicar-se o método mais simples, a eliminação de todos os valores omissos obtendo uma matriz de dados completa (caso completo) pois os resultados não serão influenciados significativamente.

Muitos programas estatísticos usam técnicas não muito satisfatórias para tratar este problema, rejeitando os registos com observações em falta (caso completo) (Little e Rubin, [24]). Pelo que se a base de dados obtida (pelo caso completo) tiver características muito diferentes da inicial (com valores omissos) este processo alterará a análise toda e as características dos resultados. Logo, uma das técnicas utilizadas é a substituição destes valores omissos por um ou vários valores, imputação simples ou múltipla, respetivamente.

Existem alguns métodos para o tratamento de valores omissos, entre os quais aqueles referidos anteriormente: o método *listwise* (caso completo), imputação simples e múltipla (Veroneze, [25]).

3.8.1 Método *Listwise*

O método *listwise* transforma o problema na situação ideal, retirando todos os valores omissos de modo a ficar apenas com os indivíduos constituídos por casos completos. Ou seja, com todos os valores das variáveis preenchidos, formando uma base de dados completa. Um dos problemas deste método é, caso a percentagem de valores omissos seja elevada e com diversas características, acaba por se obter uma base de dados que não representa a amostra total, mas apenas uma parte.

3.8.2 Imputação Simples

A imputação simples é quando cada valor omisso é substituído apenas por um único valor. Existem vários tipos de imputação única: por constantes, *hot deck* e *cold deck*. A imputação por constantes é um método em que os valores inexistentes são substituídos por um valor único em cada variável, como por exemplo 0, média, moda, mediana, ... Já na imputação *hot deck* e *cold deck* o preenchimento do valor em falta é realizado a partir dos valores observados em outros indivíduos com o mesmo atributo tendo em conta o próprio conjunto de dados ou uma fonte de dados externa, respetivamente. Um dos métodos *hot deck* preferível a muitos outros é a imputação kNN, apesar dos resultados dependerem da métrica utilizada.

Imputação kNN

A imputação kNN é um método *hot deck* pois os valores omissos são preenchidos através da análise de k vizinhos mais próximos existentes na base de dados. Ou seja, este método ao encontrar um valor omisso procura os k registos totalmente preenchidos mais parecidos/semelhantes da base de dados calculando, de seguida, uma média ou mediana dos valores presentes nos k vizinhos mais próximos encontrados de modo a obter o valor pretendido.

Esta proximidade/similaridade é representada por várias medidas de distâncias entre registos, por exemplo, a distância euclidiana. O valor ausente é calculado com base nos k vizinhos mais próximos recorrendo à média, mediana ou outra característica sumária.

O valor omisso pode ser substituído por um único valor, imputação 1NN, ou pela junção de vários valores, imputação kNN (utilizando a média, mediana, ...), onde, por omissão o valor do k no método é de 10. Este método acabou por ser utilizado na análise e tratamento de valores omissos. (Beretta e Santaniello, [26])

3.8.3 Imputação Múltipla

A imputação múltipla substitui cada valor omisso por x valores sendo $x \geq 2$, formando-se x bases de dados completas. Esta técnica engloba quatro processos: imputação, análise das bases de dados, agregação de resultados e cálculo da informação faltante. O objetivo deste método é analisar individualmente cada base de dados completa obtida e no final agregar

todos os modelos obtidos num único, resultando cada coeficiente, por exemplo, da média de todos os coeficientes obtidos anteriormente.

3.9 Métodos de Seleção de Variáveis

Para o desenvolvimento de um ótimo modelo de regressão é também necessário saber o número e quais as variáveis a utilizar, visto que quanto maior o número de variáveis menor o número de graus de liberdade de várias estatísticas e, consequentemente, maior a incerteza de alguns resultados. Sendo, por isso, aconselhável selecionar apenas as variáveis realmente importantes, ou seja, aquelas que verdadeiramente caracterizam os dados em questão e explicam a variável resposta.

É nesta seleção de variáveis explicativas que entram métodos como o *backward*, *forward* e *stepwise*. Este último resulta da junção dos dois anteriores sendo, por isso, um método mais elaborado, completo e também um dos métodos utilizados durante a análise dos dados em estudo.

– O *backward* é um método que se inicia com todas as variáveis da base de dados, variáveis a analisar, e para cada uma é calculado o valor da estatística de teste. Após o cálculo desse valor é retirada a variável com menor valor de estatística de teste e inferior a um certo limite estabelecido. Depois da eliminação dessa variável repete-se todo o processo com as variáveis ainda existentes, sendo que este só termina quando todos os valores das estatísticas de teste forem superiores ou iguais ao limite estabelecido.

– O *forward* é um método em que as variáveis são introduzidas uma a uma pela ordem de maior coeficiente de correlação entre a variável dependente e a variável a introduzir juntamente com as já introduzidas. Para cada variável a introduzir calcula-se o valor da estatística de teste, se este for inferior a um certo valor estabelecido é eliminada introduzindo-se de seguida a próxima variável a testar.

– O *stepwise*, tal como referido anteriormente resulta da junção destes dois, pois é um procedimento *forward* visto que as variáveis são introduzidas uma a uma, sendo realizada em cada passo uma avaliação para garantir que as variáveis continuam relevantes após a introdução de uma nova. Quando após a introdução de uma nova variável o valor das estatísticas de teste realizado para cada uma em particular é inferior a um certo valor estabelecido, essa variável é eliminada, considerando-se a próxima e repetindo todo o processo realizado anteriormente.

3.10 Diagnóstico de *Outliers* e Observações Influentes

A análise de resíduos é a responsável pela identificação de *outliers* e observações influentes. Algumas das medidas usadas para este diagnóstico são a *leverage*, os resíduos de Pearson estandardizados e a distância de Cook (Sarkar *et al*, [27]).

A *leverage* averigua a influência que cada observação tem no ajustamento do modelo, ou seja, se a observação é ou não influente. A *leverage* é representada por h_j e corresponde ao elemento da diagonal da matriz "chapéu" definida por $H = (X'X)^{-1}X'$. Para uma observação caso a *leverage* seja um valor próximo de 1 significa que a observação é importante no

ajustamento do modelo, caso seja próxima de 0 significa que é pouco importante. Porém, caso uma observação tenha probabilidade inferior a 0.1 ou superior a 0.9 ainda que tenha valor de *leverage* reduzido pode ser considerada como tendo influência no modelo (Norušis, [28]).

Os resíduos de Pearson estandardizados (ou também designados por "studentizados") são calculados através dos resíduos estandardizados e da *leverage*, isto é, são dados por

$$r_j = \frac{e_j^{'}}{\sqrt{1-h_j}},$$

onde $e_j^{'} = \dfrac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}}$ e têm sempre variância constante e igual a 1, sob a validade do modelo. Para amostras de grande dimensão $r_j \overset{a}{\sim} \mathcal{N}(0,1)$, pelo que 95% dos valores r_j devem estar entre -1.96 e 1.96 para $\alpha = 0.05$, visto que o quantil de ordem 0.975 de $\mathcal{N}(0,1)$, $z_{0.975} = 1.96$ e todas as observações com valores de $|r_j| > 1.96$ podem ser consideradas *outliers*. Sendo assim, conclui-se que aproximadamente 95% dos resíduos r_j devem estar entre -1.96 e 1.96.

O valor da distância de Cook é calculado usando a *leverage* e os resíduos de Pearson estandardizados e determina a variação do valor dos resíduos ao retirar uma determinada observação (j) do ajustamento do modelo, isto é, é dado por (Pregibon, [29])

$$DC_j = r_j^2 \frac{h_j}{(1-h_j)}.$$

3.11 Classificação da Variável Dependente

Após a obtenção do modelo juntamente com todos os seus parâmetros é necessário verificar se o modelo está a prever corretamente, ou seja, se quando o sujeito possui a característica em estudo o modelo prevê que ele a possui (por exemplo, quando o RN morre o modelo diz que morre fornecendo um valor de probabilidade elevado e caso não morra, o modelo fornece uma probabilidade mais baixa). Para isto é necessário o estudo da eficiência que é avaliada através da sensibilidade e especificidade do modelo.

A sensibilidade é a probabilidade do modelo prever que o sujeito possui uma determinada característica em estudo sabendo que ele realmente a possui ($P[\hat{y} = 1|y = 1]$; por exemplo, sendo a característica a morte, a probabilidade da previsão ser morte sabendo que o RN morreu), ou seja, é a percentagem de acertos nas classificações do sucesso (morte).

A especificidade é a probabilidade do modelo prever que o sujeito não possui a característica em estudo sabendo que realmente não a possui ($P[\hat{y} = 0|y = 0]$; por exemplo, a probabilidade da previsão ser não morte sabendo que o RN não morreu), ou seja, é a percentagem de acertos nas classificações de insucesso (não morte).

Estes resultados também podem ser analisados através da matriz de confusão. Esta matriz é numérica e constituída por quatro células: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP), Falsos Negativos (FN). É possível observar a matriz de confusão na tabela 3.1, onde a taxa de verdadeiros positivos corresponde à sensibilidade e a taxa de verdadeiros negativos à especificidade. Por outro lado, a taxa de falsos negativos

é a percentagem de observações presentes na classe de sucesso mas que o modelo classifica como pertencentes à classe de insucesso ($P[\hat{y} = 0|y = 1]$) e a taxa de falsos positivos é a percentagem de observações presentes na classe insucesso mas que o modelo classifica como pertencentes à classe sucesso ($P[\hat{y} = 1|y = 0]$).

	$y = 0$	$y = 1$
$\hat{y} = 0$	VN	FN
$\hat{y} = 1$	FP	VP

Tabela 3.1: Construção da matriz de confusão

Um modelo é dito com boas capacidades preditivas quando apresenta uma sensibilidade e especificidade com valores superiores a 80%. Para percentagens entre 50% a 80% diz-se que tem capacidades preditivas razoáveis e abaixo de 50% muito fracas.

Outra medida de classificação do modelo é a área sob a curva das características operacionais do recetor (ROC) discriminando os sujeitos com e sem a característica de interesse. Esta área varia entre 0 e 1 sendo que quanto mais próximo de 1 melhor é feita esta discriminação (melhor capacidade do modelo para discriminar os indivíduos que apresentam esta característica de interesse). Porém, um modelo pode possuir um valor de área sob a curva (AUC) elevado e não prever da melhor forma as probabilidades de sucesso observadas. Hosmer e Lemeshow sugerem uma classificação do poder discriminante do modelo em função do valor AUC da curva ROC, que se apresenta na tabela 3.2 (Hosmer e Lemeshow, [20]).

Área sob a curva ROC	Poder discriminante do modelo
0.5	Sem poder discriminante
]0.5, 0.7[	Discriminação fraca
[0.7, 0.8[	Discriminação aceitável
[0.8, 0.9[	Discriminação boa
≥ 0.9	Discriminação excecional

Tabela 3.2: Poder discriminante do modelo associado ao valor de AUC da curva ROC

CAPÍTULO 4

Problema em Estudo

4.1 Introdução

Para o desenvolvimento do modelo de previsão do risco de morte foi utilizada uma base de dados do recém-nascido prematuro de muito baixo peso fornecida pela SPN com informação relativa aos anos de 2005 a 2016 (1º trimestre). Nesta base de dados foram registados todos os recém-nascidos com peso inferior a 1501 gramas (independentemente da idade gestacional) ou idade gestacional menor que 32 semanas (independentemente do peso) ou ainda gémeos de gémeos que cumprissem os critérios acima.

A metodologia seguida foi a seguinte: análise e recodificação da base de dados, tratamento dos valores omissos (valores em falta), seleção das variáveis a utilizar, obtenção de diferentes modelos, escolha do modelo final e a produção de um algoritmo. Todo o trabalho foi desenvolvido usando a linguagem e ambiente R.

Inicialmente, analisou-se e recodificou-se a base de dados tendo em conta os critérios de inclusão e instruções de preenchimento de 2010 fornecidas pela SPN, nos quais foram detetados alguns problemas de qualidade de dados. De forma a diminuir o impacto destes problemas foram tomadas algumas considerações juntamente com a empresa. Não surgiram apenas estes problemas de qualidade de dados mas também a ausência de imensos valores (classificados como NA's) que tiveram de ser tratados recorrendo a um método de imputação. Seguidamente, foi necessário fazer a seleção de variáveis tendo em conta três métodos diferentes. Para cada método de seleção foi obtido um modelo, que depois dos três analisados conduziu a um modelo final que descrevesse da melhor forma o objetivo proposto. Após a obtenção do modelo foi utilizada a *package* Shiny do RStudio com o objetivo de produzir o algoritmo de modo a que o modelo pudesse ser universal e usado por todos, principalmente pelos mais interessados, os profissionais de saúde.

4.2 Caracterização da Base de Dados

A base de dados fornecida pela SPN era constituída por 12269 observações (indivíduos) e 122 variáveis. Porém, visto que o objetivo final do modelo era a previsão do risco de morte

tendo em conta apenas as variáveis desde a gravidez até à sala de partos foram somente consideradas essas. Sendo assim, a base de dados a ter em conta passou a ser constituída apenas por 47 variáveis.

Todas estas 47 variáveis foram analisadas individualmente de modo a selecionar as realmente importantes para o estudo em questão, detetando alguns erros/falhas de informação que pudessem existir.

4.2.1 Descrição das Variáveis

Das 122 variáveis disponíveis na base de dados apenas 47 variáveis eram adequadas para o problema em estudo. Dessas 47 variáveis, 28 eram de natureza qualitativa e 19 de natureza quantitativa. Sendo descritas posteriormente, tendo em conta a base de dados e o Anexo A, as variáveis que foram realmente consideradas.

A Idade da Mãe vinha expressa em anos e tinha uma média 30.5 anos e cerca de 50% das idades eram iguais ou inferiores a 31 anos (mediana = 31). O seu valor mínimo era de 13 anos e o máximo 50 anos. Existiam 211 valores 0 sem qualquer correspondência; provavelmente porque terá sido esquecido aquando do registo.

A Idade Gestacional, registada em dias aquando do nascimento, tinha uma média de 208.7 dias e mediana de 211 dias, o valor mínimo registado era 154 dias e o máximo 280 dias. Existiam 5 valores omissos.

O Peso do recém-nascido foi o primeiro obtido registado em gramas; tendo uma média de 1207, mediana de 1235, valor mínimo de 290 e máximo de 2810 gramas e uma amplitude de 2520 gramas. Esta amplitude é bastante elevada mostrando, por isso, uma grande discrepância de valores. Sendo ainda constituído por 8 valores omissos.

O Comprimento do recém-nascido, registado após o nascimento, tinha uma média 37.4 cm e mediana 38 cm, encontrando-se entre 21 e 49 cm, e tendo 1033 valores omissos.

O Perímetro Cefálico, registado ao nascer, tinha uma média de 26.8 cm e mediana de 27 cm; tendo como valor mínimo 2.8 cm e máximo 40, existindo 933 valores omissos.

O Nascimento Outborn, relativo ao local de nascimento, tomava o valor "Inborn" caso o RN tivesse nascido no hospital responsável pelo registo ou "Outborn" caso tivesse nascido fora do hospital do registo. Existiam 11489 casos "Inborn" e 734 casos "Outborn", sendo que estes últimos representavam, aproximadamente, apenas 6% dos nascimentos, ou seja a maioria dos nascimentos fora realizada no hospital responsável pelo registo. Existiam 46 valores omissos.

O Nascimento Tipo Local referia-se ao tipo de local onde nasceu. Possibilidades: "Hospital de Apoio Perinatal" (542 casos), "Hospital de Apoio Perinatal Diferenciado" (38 casos), "Hospital Privado" (1 caso), "Instituição de Saúde sem Apoio Perinatal" (7 casos), "Local Extra Hospitalar" (54 casos). A maioria dos RN nasceram no "Hospital de Apoio Perinatal" e existiam 11627 valores omissos.

O Nascimento Unidade de Saúde era constituído por 51 unidades de saúde diferentes, referenciadas apenas por números, sem re-codificação, pelo que não se conseguia saber a que unidades diziam respeito. Existiam ainda 11723 valores em branco.

O Transporte do recém-nascido podia ser por "Admissão Materna Direta", caso a mãe se tenha deslocado ao hospital por moto próprio, "Ex-Útero", caso o bebé tenha nascido noutro local e sido transferido para o hospital responsável pelo registo ou "In-Útero", se a mãe foi transferida de outro hospital com o intuito do bebé nascer no local responsável pelo registo. Existiam 7983 casos de "Admissão Materna Direta", 706 casos de "Ex-Útero", 3512 casos de "In-Útero" e ainda 68 valores omissos. Ou seja, cerca de 65.1% dos casos eram relativos à "Admissão Materna Direta".

Os Cuidados Pré-natais, cuidados obstétricos pré-natais recebidos pela mãe antes da admissão para o parto, tomavam o valor "Não" (616 casos), "Não Aplicável" (23 casos) e "Sim" (11576 casos). Existiam 54 valores omissos.

A Conceção Assistida podia ser "Sim" ou "Não", caso tenha sido ou não medicamente assistida. Existiam 1494 casos "Sim", 10670 casos "Não" e 105 valores omissos.

Os Corticoides Pré-natais, associados a ter havido ou não qualquer administração de corticoides antes do nascimento, podiam ser "Completo" (7848 casos), "Desconhecido" (30 casos), "Não" (1449 casos) e "Parcial" (2884 casos), com 58 valores omissos.

Os Corticoides Pré-natais Ciclos, relacionado com o número de ciclos de toma realizados, tomavam valores de 0 a 4 com média 1.1 e mediana 1 sendo que o 0 tinha uma frequência de 8456. Existiam 3405 valores omissos.

Os Corticoides Pré-natais Usados tomavam valores entre 1 e 4, sendo que o 1 e 2 eram os valores com maior frequência, existindo 7968 valores 1 e 2557 valores 2. Existiam ainda 1597 valores omissos, porém os números dos Corticoides Pré-natais Usados não tinham descrição associada.

As Patologias Na Gravidez podiam ser "Sim" ou "Não", caso tivesse ou não havido patologia materna durante a gravidez. Existiam 6749 casos com "Sim", 5385 com "Não" e ainda 135 valores omissos.

O Tipo de Parto podia ser "Vaginal" ou "Cesariana", existindo 8727 partos por "Cesariana", 3496 partos por "Vaginal" e 46 valores omissos. Ou seja, o parto por "Cesariana" representava cerca de 71.1% do total.

O Motivo de Parto podia ser "Espontâneo", "IVG" (Interrupção Voluntária da Gravidez), "Patologia Fetal" ou "Patologia Materna", tendo 5720, 71, 3564 e 2712 casos, respetivamente. Pelo que o parto "Espontâneo" representava cerca de 46.6% dos dados, existindo 202 valores omissos.

O Sexo podia ser "Feminino" ou "Masculino", sendo que existiam 5924 casos com o sexo Feminino e 6296 casos com o sexo Masculino, ainda três casos com o valor 95 e 43 valores omissos.

O Gemelar indicava se o RN era resultante de gestação simples ou múltipla, ou seja, se eram gêmeos ou não, existindo 4092 "Sim" e 8132 "Não", no qual cerca de 66.3% dos casos não eram gêmeos e 45 valores eram omissos. Caso a categoria da variável Gemelar fosse "Sim", o RN tinha associada uma ordem de nascimento, Gemelar Ordem, indicando o número de ordem de nascimento em questão. Este tomava valores entre 0 e 4, tendo 8179 valores omissos. O Gemelar Total, aquando da gestação múltipla era registado o número total de fetos, tomava

valores entre 1 e 4, dos quais 8179 eram valores omissos.

A Gemelar Corionicidade indicando se os fetos partilhavam ou não a mesma placenta era constituída por "Bi/Tricoriónico" (2779 casos) caso em que cada feto tem a sua própria placenta, "Desconhecido" (170 casos), "Monocoriónico" (1085 casos) caso em que todos se encontram na mesma placenta, "Não Aplicável" (8177 casos), sendo que o "Não Aplicável" representava cerca de 66.6% da totalidade, existindo 58 valores omissos.

O índice de Apgar foi registado ao 1º, 5º e 10º minutos pelo que existiam 3 variáveis relativas a esse índice, Apgar1, Apgar5 e Apgar10, respetivamente. O Apgar1 tomava valores entre 0 e 10, sendo que a sua média era de 6.6 e a mediana de 7, o valor com maior frequência era o 9 representando cerca de 24%, o 8 representava 20.7%, o 10 representava 1,3% e existiam 136 valores omissos. O Apgar 5 tomava valores entre 0 e 10, sendo que a sua média era de 8.4, a mediana 9 e o valor com maior frequência era o 9 representando cerca de 32% , o 8 representava cerca de 21.7%, o 10 representava 25.6% do total e existiam 143 valores omissos. O Apgar 10 tomava valores entre 0 e 10, sendo que a sua média era de 8.9, mediana de 9, o valor com maior frequência era o 9 representando cerca de 32.2%, o 8 representava cerca de 14.8%, o 10 representava 30.1% do total e existiam 1879 valores omissos. Concluindo-se que à medida que os minutos iam passando o índice de Apgar ia aumentando, ou seja, o RN estava a melhorar significativamente, porém o número de valores omissos aumentou.

Existiam 5 variáveis relativamente aos tipos de Ressuscitação: Ressuscitação por Adrenalina, Ressuscitação por Oxigênio, Ressuscitação por Insuflador, Ressuscitação por Entubação Et e Ressuscitação por Compressão Cardíaca. Todas eram constituídas pelas categorias "Sim", "Não" ou "Desconhecido".

A Malformação Congénita Major indicava se tinha sido ou não diagnosticada ao RN alguma malformação congénita major, podendo tomar os valores "Sim", "Não" e "Não Aplicável". Existiam 560 casos com valor "Sim", 11493 casos "Não", 99 casos "Não Aplicável" e 117 valores omissos. Concluindo-se que cerca de 93.7% dos recém-nascidos não tinham malformação congênita major.

O Surfactante Inicial, indicava se o RN recebera ou não surfactante exógeno na reanimação inicial na sala de partos. Este tomava os valores "Sim", "Não" e "Desconhecido", tendo respetivamente, 946, 11230, 26 casos, e ainda 67 valores omissos; ou seja, cerca de 91.5% dos recém-nascidos não receberam surfactante inicial.

4.2.2 Qualidade de Dados

A base de dados não estava livre de erros e tinha imensos valores omissos (descritos anteriormente em cada variável) pelo que das 47 variáveis iniciais nem todas foram importantes para a obtenção de modelos.

Análise e comparação de variáveis

Relativamente às variáveis GemelarOrdem e GemelarTotal concluiu-se que existiam 4090 casos em que os valores omissos coincidiam, não havendo nenhum caso em que a ordem ou o total estivesse preenchido e o outro não, o que revelou veracidade dos dados.

Analisando o NascimentoOutborn e o NascimentoTipoLocal notou-se que o número de dados que não eram valores omissos na variável NascimentoTipoLocal era o mesmo número de dados que eram "Outborn" na variável NascimentoOutborn. Ou seja, apenas quando o nascimento acontecia fora do local responsável pelo registo é que a variável NascimentoTipoLocal era preenchida, pois só era necessário saber o local de nascimento quando o RN não nascia na instituição do registo.

Comparando os valores omissos existentes no NascimentoTipoLocal e no NascimentoUnidadeSaude detetou-se que a variável NascimentoTipoLocal tinha 642 dados preenchidos e a variável NascimentoUnidadeSaude 546 dados preenchidos, concluindo-se que existiam 96 dados de NascimentoTipoLocal sem correspondência com a variável NascimentoUnidadeSaude. Foram então analisados os locais em que a unidade de saúde não estava preenchida: um desses locais era o "Local Extra Hospitalar", que fazia todo o sentido não ter correspondência com a unidade pois era fora do hospital, sendo que podia ter sido em qualquer outro lugar: na ambulância, em casa, não se sabe. Retirando esses casos dos 96 inicialmente obtidos restaram 42 que deveriam ter realmente correspondência da unidade de saúde e não tinham. Estes tipos de locais de nascimento eram respetivamente: "Hospital de Apoio Perinatal" (36 casos), "Hospital de Apoio Perinatal Diferenciado" (2 casos), "Hospital Privado" (1 caso), "Instituição de Saúde sem Apoio Perinatal" (3 casos). Ou seja, o maior caso em que não sabíamos a Unidade de Saúde era quando o Hospital era de Apoio Perinatal.

Analisando conjuntamente os CorticoidesPrenatais, CorticoidesPrenataisCiclos e CorticoidesPrenataisUsado concluiu-se que os valores omissos existentes na variável CorticoidesPrenatais (58 NA's) eram comuns às variáveis correspondentes: CorticoidesPrenataisCiclos e CorticoidesPrenataisUsado, o que relativamente à qualidade dos dados faz todo o sentido.

Não foram consideradas no modelo inicial as variáveis: CorticoidesPrenataisCiclos, CorticoidesPrenataisUsado, GemelarOrdem, GemelarTotal, ObitoAutopsia, ObitoCausa, ObitoAbtsençãoCuidadosterapeuticos, IdadeObitoEmDias e IdadeObitoEmHoras. As variáveis CorticoidesPrenataisCiclos e CorticoidesPrenataisUsado não foram consideradas devido à grande quantidade de valores omissos, e ao facto de serem quantitativas pelo que teriam necessariamente de ser substituídas por um valor numérico de modo a serem utilizadas pelo modelo. Porém, mesmo que o NA significasse ausência de valor ao colocar estes NA's a 0 iria alterar os dados visto que calcular o valor da média com zeros produziria resultados diferentes de fazer esta média sem valores nas mesmas variáveis. Em relação às variáveis GemelarOrdem e GemelarTotal os valores omissos representavam cerca de 66.7% dos dados totais e visto que influenciar estes valores alteraria imenso a realidade dos dados, acabaram por não ser utilizadas. Também as variáveis ObitoAbtsençãoCuidadosterapeuticos, ObitoAutopsia, ObitoCausa, IdadeObitoEmDias e IdadeObitoEmHoras não foram consideradas pelo simples fato do objetivo ser apenas prever o óbito.

Após todas estas alterações explicitadas anteriormente, a base de dados final considerada era então formada por 12220 observações e 28 variáveis.

Realizando uma análise exploratória inicial à base de dados obtiveram-se alguns resultados

fora do normal, não existentes na recodificação fornecida presente no Anexo B. A tabela 4.1 mostra esses resultados encontrados e a respetiva recodificação efetuada. Da análise da tabela concluiu-se que no total existiam 1341 valores absurdos, dos quais cerca de 72.8% eram relativos ao Apgar10. Verificou-se ainda que os valores 0 presentes na idade da mãe poderiam ter sido devidos a esquecimento aquando do registo ou falta de informação, os 99 do Apgar a um erro de escrita (pois o -999 encontrava-se na recodificação). Supôs-se o mesmo relativamente ao Perímetro Cefálico e aos Cuidados Pré-natais, visto que na recofidicação havia o -999 e o 999 respetivamente, onde o -999 significava "Não Aplicável"; porém sendo uma variável numérica teve de se substituir por NA. Já na variável sexo existiam 3 valores 95 que sem recodificação correspondente foram também substituídos por NA. Estes NA's foram de seguida modificados juntamente com todos os outros já existentes na base de dados tendo em conta um método de tratamento apropriado de valores omissos.

Variáveis	Valor	Frequência	Recodificação
Idade da Mãe	0	211	NA
Sexo	95	3	NA
Perímetro Cefálico	9999	1	-999 = NA
Cuidados Pré-natais	9	23	999 = Não Aplicável
Apgar1	99	60	-999 = NA
Apgar5	99	67	-999 = NA
Apgar10	99	976	-999 = NA

Tabela 4.1: Recodificação ausente da base de dados

Criação de novas variáveis

Para uma análise mais viável, simples e concreta foram criadas três variáveis: Ressuscitacao, Obito e RefP10:

− a variável Ressuscitacao com as categorias "Sim" e "Não" foi criada tendo em conta os "Sim" dos cinco tipos de ressuscitação (Oxigênio, Insuflador, Entubação Et, Compressão Cardíaca e Adrenalina), uma vez que o objetivo era saber se o RN teria sido sujeito a algum tipo de ressuscitação;

− a variável Obito com as categorias "Sim" e "Não" baseando-se nas três variáveis relativas à morte (MorteNaSalaDePartos, InternamentoDestino e TransferênciaDestino), para identificar se tinha ocorrido morte;

− a RefP10, sendo uma variável de referência relativamente ao percentil 10, indicava se o recém-nascido se encontrava acima ou baixo desse percentil relativamente ao peso, calculado com base neste e na idade gestacional.

A variável Ressuscitacao era constituída por 8055 casos "Sim" e 4214 "Não", a Obito por 1403 "Sim" e 10866 "Não". Concluiu-se que cerca de 65.7% dos recém-nascidos sofreram ressuscitação mas apenas 11.4% morreram.

4.3 Imputação dos Valores Omissos

A base de dados não estava totalmente completa existindo por isso imensos valores omissos. Visto que o modelo não conseguia calcular previsões com valores omissos nas variáveis de natureza quantitativa foi necessário recorrer a um método de imputação.

O método de imputação utilizado foi a imputação simples *hot deck*, método onde os valores omissos foram substituídos tendo em conta a base de dados inicial, mais propriamente a imputação kNN (k vizinhos mais próximos), através do preenchimento de todos os valores omissos existentes em todas as variáveis, tanto quantitativas como qualitativas.

Pelo método de imputação kNN quando era detetado um valor omisso, era feita a procura dos 11 (k=11) registos totalmente preenchidos mais análogos da base de dados; esta medida de similaridade era quantificada pela distância euclidiana. O valor ausente foi então calculado com base na mediana dos valores presentes nos 11 vizinhos mais próximos encontrados.

O valor escolhido para k foi 11 e a medida utilizada para obter o valor resultante destes 11 valores foi a mediana. A justificação para este procedimento foi o seguinte: em primeiro lugar, o valor por defeito de k utilizado pelo método era 10 e uma vez que o valor ausente seria substituído pela mediana, era preferível usar uma dimensão de amostra ímpar pois desta resultaria apenas um número inteiro. Desta forma não se corria o risco de obter números decimais, o que em algumas variáveis não seria muito favorável; por exemplo na idade da mãe, não poderia dar 36.6 anos. Por outro lado, a escolha da mediana é justificada por ser uma medida robusta, significando que é menos sensível a valores muito ou pouco elevados.

Apenas os valores omissos da variável Sexo não foram substituídos tendo sido eliminados uma vez que existiam 43 casos em branco e 3 valores omissos representando apenas 0.4% dos dados (46 em 12269); ou seja, era uma minoria e sendo assim não haveria necessidade de influenciar os resultados a nível da variável Sexo.

Após a imputação de todos os valores omissos, a partir da base de dados completa foram obtidas duas bases de dados treino e teste. A base de dados treino era constituída por 8554 indivíduos e a teste por 3666, sendo que ambas apresentavam as características da base de dados inicial. Esta construção fez-se uma vez que o modelo é treinado com 70% dos dados iniciais e validado com 30% (obtenção das capacidades preditivas).

4.4 Seleção de Variáveis

Após a análise de todas as variáveis, adição de algumas e eliminação de outras obtiveram-se as variáveis a incluir no modelo de regressão logística. A tabela 4.2 apresenta as variáveis selecionadas com a notação correspondente e respetivas categorias. Quando a variável não for qualitativa mas quantitativa a categoria encontra-se com um "−", significando que não existe.

Após esta seleção das variáveis da base de dados, procedeu-se à construção de possíveis modelos de regressão logística, usando diferentes métodos de seleção.

Visto que foram considerados três métodos diferentes (teste de Wald, análise de variância (através do teste do qui-quadrado) e *stepwise*) resultaram três modelos diferentes. Com base

em cada método de seleção de variáveis, do qual resultou um conjunto de variáveis distintas, foi proposto um modelo de regressão logística. Porém, da análise dos correspondentes *p-values* resultantes da aplicação do teste de Wald para os três modelos obtidos, para estudar a significância dos correspondentes parâmetros, constatou-se que as categorias "Desconhecido" da variável CorticoidesPrenatais e "Não Aplicável" da Malformação Congénita não eram significativas, pois apresentavam um *p-value* superior a 0.9 e um valor de erro padrão demasiado elevado. Sendo assim, visto que estas duas categorias representavam cerca de 0.7% dos resultados totais e só tinham necessidade de aparecer no inquérito caso houvesse falta de informação, estas duas categorias foram substituídas por uma categoria já existente, a categoria "Não", de modo a preservar todos os outros resultados existentes. Não se poderia prejudicar o modelo ou eliminar as variáveis em causa devido apenas a categorias que acabam por ser indispensáveis aquando do registo.

Tendo em conta as alterações anteriormente mencionadas obtiveram-se três novos modelos (modelos 1, 2 e 3) que se passam a considerar. As tabelas 4.3, 4.4 e 4.5 indicam os resultados da aplicação do teste de Wald para os três modelos obtidos para estudar a significância dos correspondentes parâmetros. Da análise dos respetivos *p-values* constatou-se que apesar de nem todas as categorias serem demasiado significativas as variáveis correspondentes eram-no e, por isso, foram todas consideradas na construção dos modelos. Por exemplo, considerando o modelo 1, obtiveram-se os seguintes *p-values* consideravelmente elevados: 0.5905, 0.6698 e 0.6092 correspondentes à categoria Parcial da variável Corticoides Pré-natais e às categorias IVG e Patologia Materna do Motivo de Parto. Porém, apesar destas categorias terem um *p-value* demasiado elevado, não se poderia abdicar da informação da variável em questão (Corticoides Pré-natais e Motivo de Parto), obtendo-se conclusões semelhantes para os dois outros modelos (modelo 2 e 3).

Variáveis	Notação	Categorias
Idade da Mãe	MaeIdade	–
Idade Gestacional	IdadeGestacional	–
Peso	NascimentoPeso	–
Comprimento	NascimentoComprimento	–
Perímetro Cefálico	NascimentoPerimetroCefalico	–
Outborn	NascimentoOutborn	1 - Outborn 2 - Inborn
Tipo de Local	NascimentoTipoLocal	1 - Hospital de Apoio Perinatal Diferenciado 2 - Hospital de Apoio Perinatal 3 - Instituição de Saúde sem Apoio Perinatal 4 - Local Extra Hospitalar 5 - Hospital Privado
Unidade de Saúde	NascimentoUnidadeSaude	Unidade 1 . . . Unidade 56
Transporte	Transporte	1 - In-Útero 2 - Ex-Útero 3 - Admissão Materna Direta
Cuidados Prenatais	CuidadosPrenatais	1 - Sim 2 - Não 999 - Não Aplicável
Conceção Assistida	ConcepcaoAssistida	1 - Sim 2 - Não 999 - Não Aplicável
Corticoides Pré-natais	CorticoidesPrenatais	1 - Não 2 - Parcial 3 - Completo 4 - Desconhecido 999 - Não Aplicável
Patologias na Gravidez	PatologiasNaGravidez	1 - Sim 2 - Não 999 - Não Aplicável
Tipo de Parto	TipoDeParto	1 - Vaginal 2 - Cesariana
Motivo do Parto	MotivoDoParto	1 - Espontâneo 2 - Patologia Materna 3 - Patologia Fetal 4 - IVG
Sexo	Sexo	1 - Masculino 2 - Feminino
Gemelar	Gemelar	1 - Sim 2 - Não 999 - Não Aplicável
Apgar1	Apgar1	–
Apgar5	Apgar5	–
Apgar10	Apgar10	–
Ressuscitação	Ressuscitacao	1 - Sim 2 - Não
Malformação Congénita	MalformacaoCongenitaMajor	1 - Sim 2 - Não 999 - Não Aplicável
Percentil 10	RefP10	1 - Inferior ao P10 0 - Superior ou igual ao P10

Tabela 4.2: Descrição das variáveis a considerar na obtenção dos modelos

Variáveis Independentes	β	S.E.	Valor do Teste	Significância (*p-value*)
Constante	14.7386	0.8609	17.1210	$< 2\text{x}10^{-16}$
IdadeGestacional	-0.0437	0.0044	-9.9520	$< 2\text{x}10^{-16}$
NascimentoPeso	-0.0009	0.0003	-2.8800	0.0040
NascimentoComprimento	-0.1034	0.0227	-4.5450	$5.49\text{x}10^{-6}$
CorticoidesPrenataisNão	0.6174	0.1279	4.8280	$1.38\text{x}10^{-6}$
CorticoidesPrenataisParcial	-0.0549	0.1020	-0.5380	0.5905
TipoDePartoVaginal	0.3293	0.1080	3.0480	0.0023
MotivoDoPartoIVG	-0.1670	0.3917	-0.4260	0.6698
MotivoDoPartoPatologia Fetal	0.5082	0.1254	4.0540	$5.04\text{x}10^{-5}$
MotivoDoPartoPatologia Materna	-0.0643	0.1259	-0.5110	0.6092
SexoMasculino	0.4384	0.0863	5.0800	$3.78\text{x}10^{-7}$
Apgar1	-0.1184	0.0227	-5.2190	$1.80\text{x}10^{-7}$
Apgar10	-0.4164	0.0395	-10.5450	$< 2\text{x}10^{-16}$
MalformacaoCongenitaMajorSim	1.4534	0.1578	9.2130	$< 2\text{x}10^{-16}$

Tabela 4.3: Características dos coeficientes do modelo 1

Variáveis Independentes	β	S.E.	Valor do Teste	Significância (*p-value*)
Constante	12.9325	1.1023	11.7320	$< 2\text{x}10^{-16}$
IdadeGestacional	-0.0321	0.0061	-5.2260	$1.73\text{x}10^{-7}$
NascimentoPeso	-0.0015	0.0004	-3.9560	$7.61\text{x}10^{-5}$
NascimentoComprimento	-0.1071	0.0229	-4.6730	$2.96\text{x}10^{-6}$
TransporteEx-Utero	-0.2474	0.1686	-1.4670	0.1423
TransporteIn-Utero	-0.1478	0.0958	-1.5420	0.1230
CorticoidesPrenataisNão	0.6522	0.1385	4.7090	$2.49\text{x}10^{-6}$
CorticoidesPrenataisParcial	-0.0418	0.1030	-0.4060	0.6848
PatologiasNaGravidezSim	-0.1432	0.0959	-1.4920	0.1356
TipoDePartoVaginal	0.3744	0.1099	3.4070	0.0007
MotivoDoPartoIVG	-0.1409	0.3953	-0.3570	0.7215
MotivoDoPartoPatologia Fetal	0.6009	0.1293	4.6460	$3.38\text{x}10^{-6}$
MotivoDoPartoPatologia Materna	0.1340	0.1390	0.9640	0.3349
SexoMasculino	0.4437	0.0866	5.1240	$2.99\text{x}10^{-7}$
GemelarSim	0.2355	0.0955	2.4640	0.0137
Apgar1	-0.1234	0.0228	-5.4140	$6.17\text{x}10^{-8}$
Apgar10	-0.4140	0.0396	-10.4430	$< 2\text{x}10^{-16}$
MalformacaoCongenitaMajorSim	1.4630	0.1583	9.2420	$< 2\text{x}10^{-16}$
RefP10Superior ou igual ao P10	0.3958	0.1484	2.6680	0.0076

Tabela 4.4: Características dos coeficientes do modelo 2

Variáveis Independentes	β	S.E.	Valor do Teste	Significância (*p-value*)
Constante	15.3392	0.8407	18.2460	$< 2\text{x}10^{-16}$
IdadeGestacional	-0.0463	0.0043	-10.6860	$< 2\text{x}10^{-16}$
NascimentoPeso	-0.0008	0.0003	-2.6690	0.0076
NascimentoComprimento	-0.1040	0.0227	-4.5780	$4.68\text{x}10^{-6}$
CorticoidesPrenataisNão	0.6470	0.1271	5.0900	$3.58\text{x}10^{-7}$
CorticoidesPrenataisParcial	-0.0363	0.1017	-0.3570	0.7209
MotivoDoPartoIVG	-0.1066	0.3895	-0.2740	0.7843
MotivoDoPartoPatologia Fetal	0.3705	0.1156	3.2040	0.0014
MotivoDoPartoPatologia Materna	-0.1965	0.1176	-1.6720	0.0946
SexoMasculino	0.4270	0.0861	4.9600	$7.06\text{x}10^{-7}$
Apgar1	-0.0978	0.0263	-3.7150	0.0002
Apgar5	-0.0461	0.0425	-1.0840	0.2785
Apgar10	-0.3851	0.0491	-7.8430	$4.38\text{x}10^{-15}$
MalformacaoCongenitaMajorSim	1.4694	0.1580	9.3020	$< 2\text{x}10^{-16}$

Tabela 4.5: Características dos coeficientes do modelo 3

4.5 Avaliação da Qualidade dos Modelos e das Capacidades Preditivas

Seguidamente pretendeu-se fazer um estudo comparativo da qualidade dos modelos e das suas capacidades preditivas.

A tabela 4.6 apresenta valores de algumas características usualmente consideradas para avaliar a qualidade dos modelos ajustados. Assim, calculou-se o valor do critério de informação de Akaike (AIC), dado por (Hox, [30])

$$AIC = -2L(\beta, \theta) + 2p,$$

onde L é a log-verosimilhança e o p o número de variáveis independentes do modelo (preditores), para se analisar qual o modelo que seria melhor ajustado. Verificou-se que, em termos deste indicador, o modelo 3 é o que é pior modelado. No entanto, as diferenças não são significativas.

Em relação ao teste de Hosmer & Lemeshow, teste de ajustamento do modelo aos dados, verificou-se um *p-value* de destaque relativamente ao modelo 2, mostrando que se ajusta bem melhor aos dados que os outros dois. Relativamente aos resíduos de Pearson "studentizados" a percentagem de valores entre -1.96 e 1.96 é superior a 95% em todos os casos, sendo um bocadinho superior no último modelo (modelo 3), a média encontrava-se próxima de 0, porém a variância (valor do desvio padrão ao quadrado) estava bastante desviada de 1.

Relativamente às características que avaliam as capacidades preditivas dos modelos (curva ROC, precisão, sensibilidade e especificidade) concluiu-se que têm todos um poder discriminante excecional (área abaixo da curva ROC acima de 0.9), tendo ainda bons valores para a precisão, sensibilidade e especificidade (acima de 80% sendo modelos com boas capacidades preditivas). Na figura 4.1 estão representadas as curvas ROC dos três modelos e os valores do *cutoff* (valores coloridos entre 0 e 1, a partir dos quais a probabilidade é considerada sucesso). Este valor é conseguido com base nos melhores valores para a sensibilidade e especificidade, sendo que nestes casos o *cutoff* foi de 0.13, 0.13 e 0.12. Porém, apesar de ser um valor baixo é o valor que melhores capacidades preditivas apresenta para os modelos.

Globalmente e como consequência de uma análise balanceada nos resultados anteriormente discutidos concluiu-se que os três modelos eram muitos semelhantes. Sendo assim, tendo em conta o número de variáveis (princípio da parcimônia) e o valor do (*p-value* do teste de Hosmer & Lemeshow) o modelo selecionado de entre os três foi o modelo 1, pois todos os valores das características tidas em conta são intermédios e o número de variáveis não é demasiado elevado e são variáveis que tal como referenciado no início do relatório contribuem para a mortalidade de recém-nascidos prematuros e de muito baixo peso.

	Modelo 1	Modelo 2	Modelo 3
Variáveis	IdadeGestacional NascimentoPeso NascimentoComprimento CorticoidesPrenatais TipoDeParto MotivoDoParto Sexo Apgar1 Apgar10 MalformacaoCongenita	IdadeGestacional NascimentoPeso NascimentoComprimento Transporte CorticoidesPrenatais PatologiasNaGravidez TipoDeParto MotivoDoParto Sexo Gemelar Apgar1 Apgar10 MalformacaoCongenita RefP10	IdadeGestacional NascimentoPeso NascimentoComprimento CorticoidesPrenatais MotivoDoParto Sexo Apgar1 Apgar5 Apgar10 MalformacaoCongenita
AIC	3872.2	3862	3880.3
AUC	0.9050	0.9025	0.9058
Precisão	0.8347	0.8317	0.8202
Sensibilidade	0.8350	0.8331	0.8177
Especificidade	0.8325	0.8208	0.8396
Teste de Hosmer & Lemeshow χ^2 df *p-value*	 6.4368 8 0.5984	 5.6191 8 0.6898	 6.5663 8 0.5841
Resíduos de Pearson "Studentizados" Média Desvio Padrão -1.96 e 1.96	 -0.1202 0.6609 97.26%	 -0.1197 0.6597 97.22%	 -0.1200 0.6617 97.38%

Tabela 4.6: Análise da qualidade dos três modelos de regressão logística obtidos

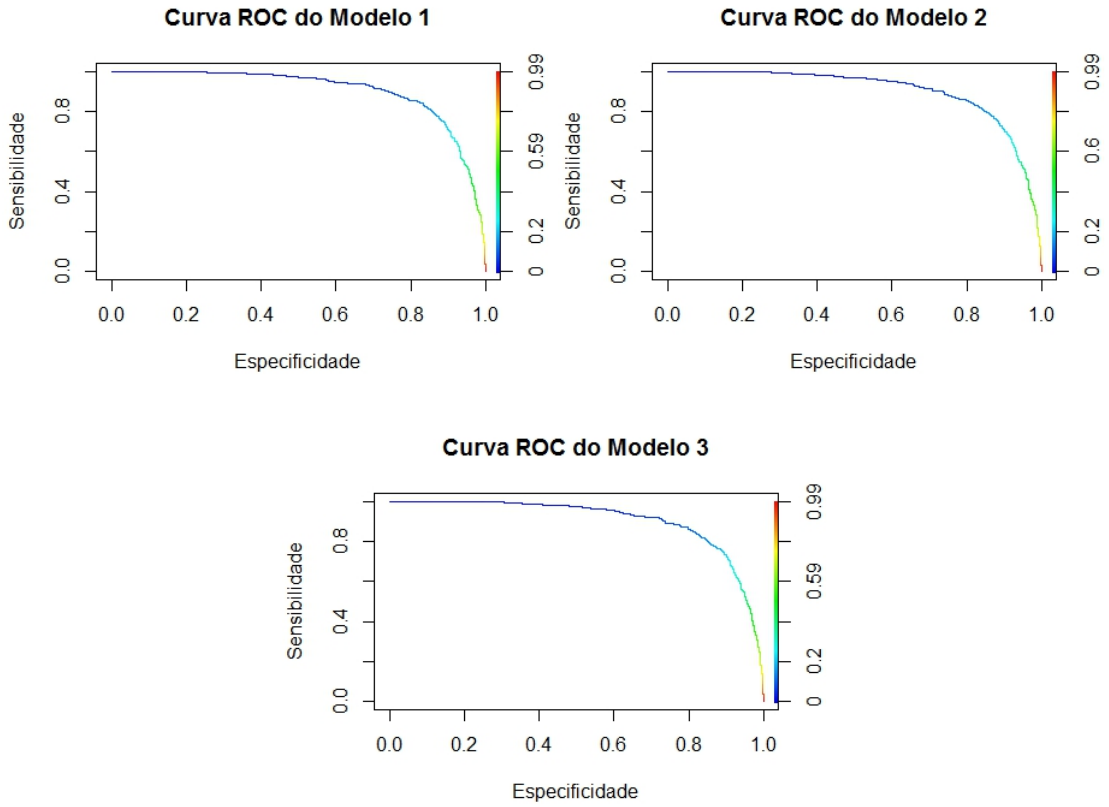

Figura 4.1: Curvas ROC dos três modelos iniciais 1, 2 e 3

4.6 Análise da Existência de Possíveis *Outliers* (Observações Influentes)

Após a seleção do modelo, o objetivo foi saber se era possível melhorá-lo, de modo a obter a partir desse um modelo ainda melhor. Para isso foram utilizados alguns processos de deteção de *outliers* ou observações influentes, a *leverage* e a distância de Cook. Analisaram-se assim estas duas medidas de maneira a detetar e eliminar possíveis *outliers* que pudessem existir.

Relativamente à *leverage*, inicialmente calculou-se o seu valor (h_j) para cada indivíduo (j); se $h_j > \frac{2(k+1)}{n}$ a observação j era considerada *outlier*, com k a representar o número de variáveis independentes e n a dimensão da amostra. Ao retirar todas as observações que se encontravam acima desse valor $\left(\frac{2(10+1)}{8554}\right)$ foi retirada a categoria "IVG" da variável MotivoDoParto, concluindo-se que nem todas essas observações eram influentes, pois o modelo necessitava delas. Sendo assim, numa segunda análise realizada através da observação gráfica da figura 4.2 dos valores da *leverage* concluiu-se que os valores de h acima de 0.020 se destacavam, porém ao retirar todas estas observações, todos os indivíduos considerados possíveis *outliers* possuíam uma característica em comum, que era também a categoria "IVG" da variável MotivoDoParto. Ou seja, após a eliminação destes valores o modelo não conseguia fazer previsões pois faltava uma categoria da variável MotivoDoParto. Tal deveu-se ao facto de na base de dados inicial apenas existirem 71 casos de "MotivoDoParto Desconhecido", sendo a base de dados treino constituída por 57 e teste por 24. Assim, ao retirar os casos de *outliers* retiravam-se todos os casos "IVG" e o modelo não conseguia fazer previsões pois eliminando esses indivíduos a categoria desaparecia da base de dados treino, e como o modelo foi treinado sem essa categoria, ao fazer previsões com essa categoria na base de dados teste, essa era considerada uma nova categoria, o que não era desejável. Foi ainda retirado o indivíduo com o valor de h acima de 0.025 (um único valor que sobressaía) e o modelo também não melhorou visto que era apenas uma observação no meio de tantas. Não tendo sido, por isso, eliminados do modelo os indivíduos com os valores de h referidos e este método acabou por não ter qualquer influência no melhoramento do modelo, uma vez que não foi utilizado.

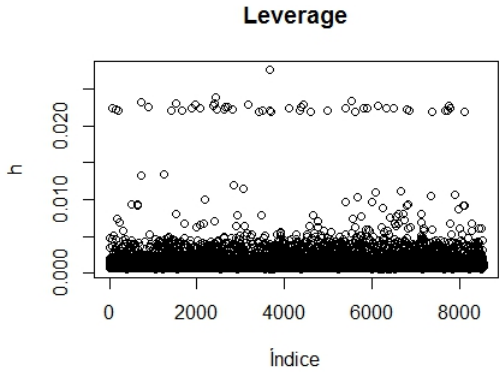

Figura 4.2: Gráfico dos valores de *leverage* do modelo 1

Relativamente à distância de Cook, esta por norma deve ser inferior a 1 de modo a não ter observações influentes. No entanto, ao calcular a distância de Cook verificou-se que eram todas inferiores a 1, o que pode ser visível através da figura 4.3, significando que não existiam observações influentes. Porém, notou-se ainda que ao retirar determinadas observações o ajustamento do modelo aos dados melhorava para um certo valor de distância de Cook, ou seja, o *p-value* do teste de Hosmer & Lemeshow passava de 0.5984 para um valor superior (0.7156), o que é possível verificar através da análise da tabela 4.7. Nesta tabela estão representados três valores de distâncias de Cook testados, juntamente com o AIC do modelo, área abaixo da curva ROC e o valor do teste de Hosmer & Lemeshow correspondentes, concluindo-se que eram aproximadamente iguais todos estes valores sendo o *p-value* de h=0.007 o valor que mais se destacava. Logo, retirando os indivíduos com distâncias de Cook superiores a 0.007 o ajustamento do modelo aos dados melhorava significativamente e o valor do AIC também (passou de 3872.2 para 3841), sendo que o valor da área abaixo da curva ROC não teve grande alteração. Pelo que se eliminaram esses indivíduos da base de dados inicial, obtendo-se um novo modelo, sendo esse o Modelo Final a considerar na previsão do risco de morte e todas as características apresentadas na tabela 4.8. Analisando este modelo através desta tabela concluiu-se, tal como dito anteriormente que existe um bom ajustamento do modelo aos dados (*p-value* = 0.7156), a percentagem de resíduos de Pearson "studentizados" com valores entre -1.96 e 1.96 é superior a 95% sendo de 97.4%, o valor da área sob a curva ROC encontra-se acima de 0.9 indicando um poder discriminante excecional do modelo e a sensibilidade e especificidade acima de 80% sendo um modelo com boas capacidades preditivas.

Após a obtenção de todas as características do Modelo Final faltava saber o valor dos parâmetros associados a cada variável (valores de β). Encontram-se apresentados na tabela 4.9, juntamente com o valor do erro padrão dos parâmetros, a sua significância resultante da aplicação do teste de Wald e o valor de e^{β}. Este último é o valor utilizado para calcular a probabilidade de morte dos RN.

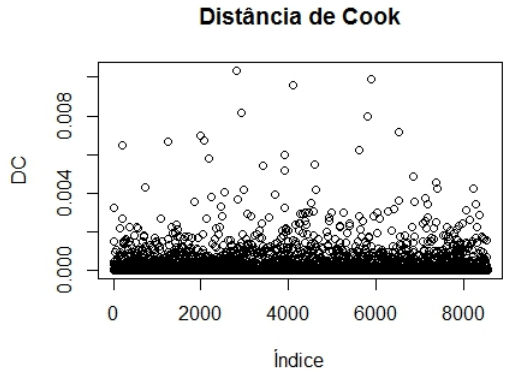

Figura 4.3: Gráfico dos valores de distância de Cook do modelo 1

Valor da Distância de Cook	AIC	AUC	Teste Hosmer-Lemeshow (*p-value*)
≤ 0.006	3813.5	0.9045	0.5174
≤ 0.007	3841	0.9047	0.7156
≤ 0.008	3850	0.9048	0.5111

Tabela 4.7: Características do modelo 1 associadas a cada valor da distância de Cook

	Modelo Final
AIC	3841
AUC	0.9047
Precisão	0.8322
Sensibilidade	0.8313
Especificidade	0.8396
Teste de Hosmer & Lemeshow	
χ^2	5.3861
df	8
p-value	0.7156
Resíduos de Pearson "Studentizados"	
Média	-0.1194
Desvio Padrão	0.6585
-1.96 e 1.96	97.31%

Tabela 4.8: Características do modelo final (modelo 1 sem *outliers*)

Variáveis Independentes	β	S.E.	Valor do Teste	Significância (*p-value*)	$Exp(\beta)$
Constante	14.7873	0.8660	17.0750	$< 2\text{x}10^{-16}$	2642737
IdadeGestacional	-0.0446	0.0044	-10.0590	$< 2\text{x}10^{-16}$	0.9564
NascimentoPeso	-0.0009	0.0003	-3.0610	0.0022	0.9991
NascimentoComprimento	-0.0991	0.0228	-4.3410	$1.42\text{x}10^{-5}$	0.9056
CorticoidesPrenataisNão	0.6401	0.1284	4.9840	$6.22\text{x}10^{-7}$	1.8966
CorticoidesPrenataisParcial	-0.0590	0.1025	-0.5750	0.5650	0.9427
TipoDePartoVaginal	0.3216	0.1086	2.9620	0.0031	1.3793
MotivoDoPartoIVG	-0.4884	0.4270	-1.1440	0.2528	0.6136
MotivoDoPartoPatologia Fetal	0.5263	0.1258	4.1830	$2.88\text{x}10^{-5}$	1.6926
MotivoDoPartoPatologia Materna	-0.0567	0.1262	-0.4490	0.6534	0.9449
SexoMasculino	0.4559	0.0867	5.2570	$1.46\text{x}10^{-7}$	1.5776
Apgar1	-0.1183	0.0228	-5.1930	$2.07\text{x}10^{-7}$	0.8885
Apgar10	-0.4165	0.0396	-10.5240	$< 2\text{x}10^{-16}$	0.6593
MalformacaoCongenita MajorSim	1.4220	0.1590	8.9450	$< 2\text{x}10^{-16}$	1.4220

Tabela 4.9: Características dos coeficientes do modelo final de regressão logística

4.7 Aplicação do Algoritmo Usando o Shiny

Tão importante como ter o modelo a funcionar e a dar probabilidades corretas era a sua fácil utilização por parte de quem necessita de o usar todos os dias, tais como, médicos e enfermeiros.

Daí surgiu a necessidade de usar o Shiny (Sistema para o desenvolvimento de aplicações *web*) em conjunto com o R. Permitindo assim a utilização do modelo por todos os que necessitam de o usar mesmo não percebendo o método utilizado para o cálculo da probabilidade, visto que o objetivo dos profissionais de saúde não é dominar o método mas interpretar o resultado e agir de diferentes formas perante ele.

Foi então criada uma aplicação *web*, apresentada na figura 4.4, em que os valores das probabilidades são calculados em tempo real para cada RN e a implementação é possível em qualquer dispositivo: telemóvel, tablet, ... Esta aplicação era constituída por um formulário com as variáveis explicativas do modelo, sendo estas a Idade Gestacional (em dias), o Peso (em gramas), o Comprimento (em cm), os Corticoides Pré-natais (Não/ Parcial/ Sim), o Tipo de Parto (Vaginal/ Cesariana), o Motivo do Parto (Espontâneo/ IVG/ Patologia Fetal/ Patologia Materna) o Sexo (Feminino/ Masculino), o Apgar1 (entre 0 e 10), o Apgar10 (entre 0 e 10) e a Malformação Congénita (Sim/ Não). Após o preenchimento do formulário ao clicar no botão "Cálculo" era apresentado um valor de probabilidade de morte juntamente com a cor associada ao risco. Sendo que a cor era diretamente proporcional ao risco, existiam três possibilidades de cores relativamente à probabilidade de morte, para valores de probabilidade de 0% a 49% o risco é reduzido sendo representado por uma cor verde, de 50% a 74% o risco é moderado sendo representado por uma cor amarela e de 75% a 100% o risco é elevado sendo representado pela cor vermelha, estes pressupostos foram tomados em conjunto com a empresa. A imagem utilizada na aplicação web foi retirada de (Frutuoso, [31]).

Figura 4.4: Apresentação da aplicação do algoritmo usando o Shiny para a previsão do risco de morte

É ainda importante referir que o preenchimento das primeiras três variáveis (Idade Gestacional, Peso e Comprimento) será realizado de forma autónoma pelo utilizador sendo que aquando da utilização de casas decimais estas deverão ser referenciadas com ponto (.) em vez de vírgula (,) de modo ao modelo calcular. É ainda enviada uma mensagem de erro aquando do preenchimento destas variáveis com valores negativos ou valor zero, indicando a variável em questão e ausência do valor do cálculo da probabilidade com o NA. Além destas, não foi imposta mais nenhuma restrição aos valores destas variáveis, visto que a qualquer momento pode nascer um bebé com valores destes fora do normal, não pondo em causa a utilidade do programa. Aquando da utilização de letras ou vírgulas nestas variáveis o valor de probabilidade também é NA, ou seja, não consegue calcular, visto que o resultado inserido é absurdo para o programa.

A variável Sexo tem duas categorias (Feminino e Masculino), os Corticoides Pré-natais têm três categorias (Não, Parcial e Sim), o Tipo de Parto tem duas categorias (Vaginal e Cesariana), o Motivo do Parto tem quatro categorias (Espontâneo, IVG, Patologia Fetal e Patologia Materna) e a Malformação Congénita tem duas categorias (Sim e Não). Estas variáveis são variáveis categóricas que podem ser acedidas carregando na seta e selecionando a opção em causa. Já o Apgar1 e o Apgar10 são variáveis numéricas que tomam valores entre 0 e 10, sendo obtidos através da utilização de duas setas que somam ou retiram uma unidade, carregando na seta de cima ou de baixo, respetivamente.

Foram ainda calculadas três previsões de risco de morte de modo ao utilizador perceber como funciona realmente esta aplicação e a influência que diversas variáveis têm na probabilidade de morte dos RN, sendo possível observar através da figura 4.5. Esta figura mostra três exemplos de recém-nascidos com características diferentes, pelo que dependendo das características a probabilidade de morte variará. Sendo que o primeiro indivíduo terá uma probabilidade de 8.9%, ou seja, tem um risco reduzido pelo que aparece a cor verde. Já o segundo com algumas características semelhantes ao primeiro e valores de Idade Gestacional, Peso, Comprimento e Apgar10 mais baixos apresentou uma probabilidade de morte mais elevada, passando de 8.9% para 54.7% e de um risco reduzido para um risco moderado, com cor amarela. O terceiro indivíduo comparativamente com o segundo difere apenas no motivo de parto tendo sido Patologia Fetal, no sexo Masculino e existência de malformação, tendo uma probabilidade de morte elevadíssima, passando de 54.7% para 93% com um risco elevado de morte, com cor vermelha.

Numa Unidade de Cuidados Intensivos Neonatais após o cálculo destes três diferentes diagnósticos apresentados na figura 4.5 o profissional de saúde iria agir de diferente forma. Sendo que o recém-nascido com maior urgência de vigilância e atuação seria o último, com probabilidade de morte mais elevada e, de seguida, o segundo com risco moderado mas que após alguns procedimentos poderia diminuir e permitir ao recém-nascido continuar a viver.

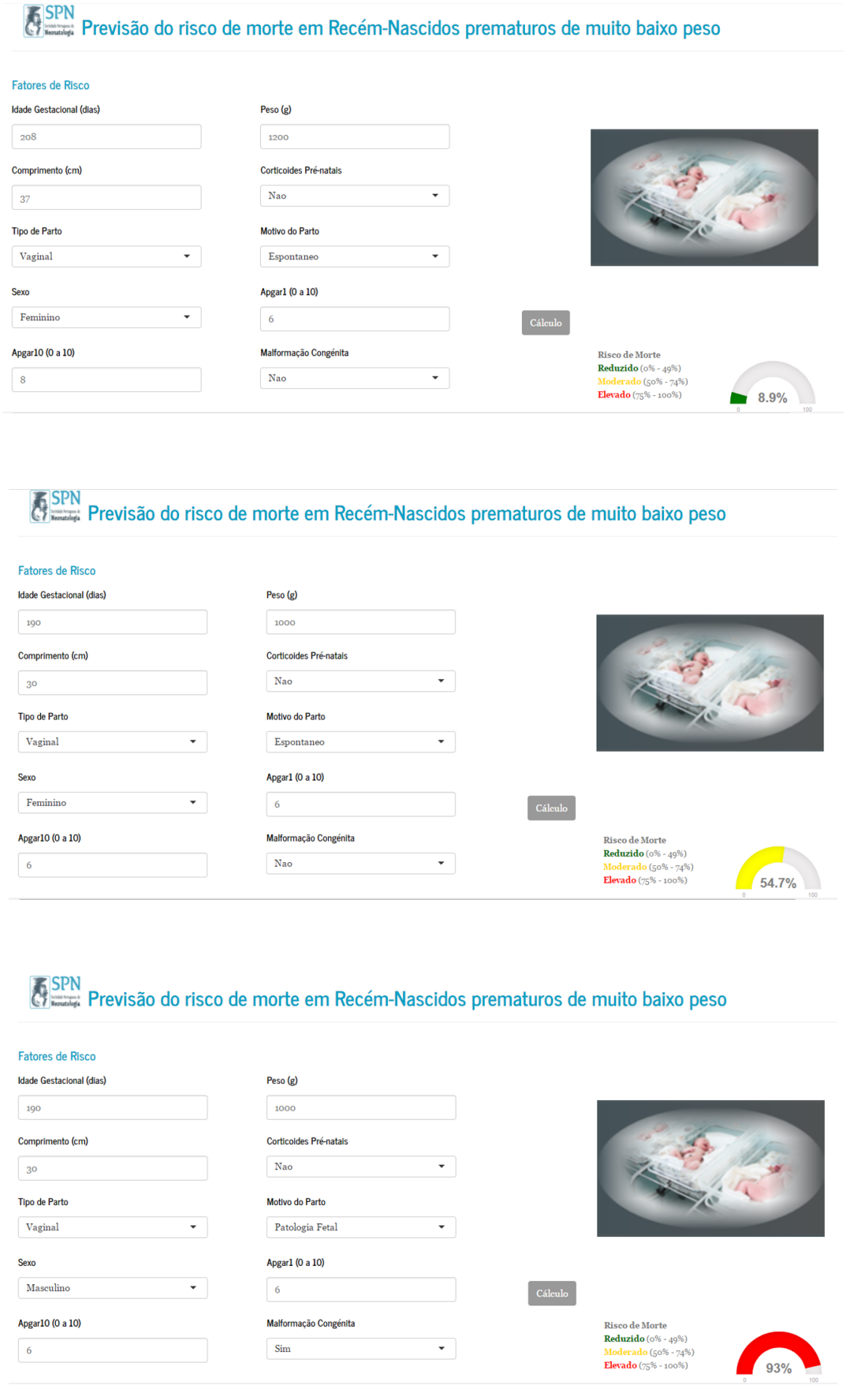

Figura 4.5: Exemplos de diferentes previsões do risco de morte utilizando o Shiny

Verificou-se com a utilização do Shiny e através destes três últimos exemplos que um reduzido valor de Idade Gestacional, Comprimento e Apgar, a ausência da toma de Corticoides Pré-natais, o tipo de parto Vaginal, o motivo de parto Patologia fetal e a existência de malformações juntamente com o sexo Masculino aumentam a probabilidade de morte do RN, tal como concluído inicialmente aquando do estudo de artigos e teoria sobre Neonatologia. Pelo que é necessário a mãe ter uma gravidez vigiada e acompanhada de um bom profissional de saúde, de modo a evitar algumas das características referidas anteriormente.

CAPÍTULO 5

Conclusão

O número de nascimentos prematuros em Portugal tem vindo a aumentar, uma razão que contribui para esse facto é as mulheres terem o seu primeiro filho cada vez mais tarde. Por isso, faz todo o sentido implementar e redobrar mecanismos de vigilância para evitar a morte destes recém-nascidos. Sendo necessário atuar, a SPN tinha em vista a implementação de um modelo preditivo com o objetivo de prever o risco de morte de recém-nascidos prematuros de muito baixo peso. Esta previsão pode auxiliar a que os profissionais de saúde estejam mais vigilantes aos recém-nascidos e perceberem quais os que têm um maior risco de morte, atuando sobre eles de modo a tentar minimizar essa ocorrência.

O modelo preditivo utilizado foi o modelo de regressão logística múltipla. O modelo obtido fornece a probabilidade do risco de morte do recém-nascido tendo em conta dez características: a Idade Gestacional (em dias), o Peso (em gramas), o Comprimento (em cm), a toma ou não de Corticoides Pré-natais, o Tipo de Parto, o Motivo do Parto, o Sexo, o Apgar1 (entre 0 e 10), o Apgar10 (entre 0 e 10) e a existência ou não de Malformação Congénita.

O modelo foi obtido com base em dados reais fornecidos pela SPN. É de salientar que este trabalho, por modelar uma situação real de grande complexidade, teve algum grau de dificuldade. Pois, na maioria das vezes, os dados reais não satisfazem os pressupostos teóricos, e pela sua elevada dimensão levantam outro tipo de problemas que ultrapassa largamente a mera ilustração de modelos teóricos, a que nós, alunos, estamos habituados a tratar. Por outro lado, dado a natureza do problema e a sua delicadeza por envolver vidas humanas, acresce uma maior responsabilidade nas decisões a tomar e no controlo do erro. É de registar que os resultados obtidos para o modelo foram bastante satisfatórios tendo em vista que o valor sob a curva ROC era de aproximadamente 0.90, valores de sensibilidade e especificidade acima de 80% e o ajustamento do modelo aos dados também era bom (teste de Hosmer & Lemeshow, *p-value*=0.7156).

Além do desenvolvimento do modelo foi utilizado o Shiny juntamente com o R de modo a proporcionar a utilização do modelo por qualquer pessoa e ainda facilitar a sua interpretação evidenciando o seu aspeto gráfico. Esta previsão é calculada em tempo real para o bebé em

questão, baseando-se apenas nas suas características. Por outro lado é ainda possível a sua implementação em qualquer dispositivo, tal como computador, telemóvel, tablet, ... de modo a poder ser usado em qualquer local.

O estágio numa empresa com partilha de conhecimento diverso entre profissionais permite ir além das expetativas e querer sempre mais. O conhecimento partilhado entre todos e a vontade de ajudar conduziu não só a uma probabilidade encontrada pelo modelo mas também à apresentação da mesma. E um objetivo de um estágio passa mesmo por aí, aprender e preparar para a vida real ajudando não só a adquirir as ferramentas em questão mas também, mais importante que isso, a querer adquiri-las e superar todas as dificuldades, ultrapassando metas superiores aos objetivos iniciais do aluno.

Desafios para o futuro:

− o cálculo da probabilidade de morte desde a gravidez após a sala de partos (tendo em conta todo o resto), pois este modelo tem em conta apenas as variáveis conseguidas até dez minutos após o nascimento do bebé;

− a determinação de variáveis decisivas relativamente à probabilidade de morte em que os médicos podem interferir de modo a diminuir essa probabilidade. Exemplificando, o modelo pode fornecer uma probabilidade mas caso seja feito algo essa probabilidade poderá diminuir, como por exemplo, a reanimação na sala de partos. Fazendo com que este modelo estático - pois para cada RN calcula apenas um valor de probabilidade - se transforme num modelo dinâmico dependente dos processos que se pudessem fazer após os dez minutos de vida do RN, a probabilidade de morte calculada inicialmente poderá diminuir drasticamente;

− o uso de outros métodos de seleção de variáveis, em particular metodologias robustas envolvendo, por exemplo o método de mínimos quadrados aparados.

Concluindo, o modelo não teve em vista apenas o cálculo da probabilidade de morte só por uma questão de curiosidade e conhecimento aquando do nascimento, mas também colocar os profissionais de saúde ainda mais atentos ao estado dos recém-nascidos prematuros. Pois, além de todos serem prematuros e com baixo peso, tendo uma probabilidade de morte superior aos RN de termo, uns terão uma probabilidade maior que outros. Sendo, consequentemente, necessário atuar mais rapidamente nos que têm um maior risco de morte, pelo que é fundamental conseguir detetá-lo, o que passa a ser possível com a realização deste trabalho.

Referências

[1] W. Grobman e D. Stamilio, «Methods of clinical prediction», *Am J Obstet Gynecol*, vol. 194, pp. 888–94, 2006.

[2] S. T. Adams e S. H. Leveson, «Clinical prediction rules», *Bmj*, vol. 344, n.º 1, pp. d8312–8319, 2012.

[3] D. D. Neves e R. M. Dias, «Testes diagnósticos 4 : algumas regras de predição clínica», *Pulmão RJ*, vol. 13, pp. 45–48, 2004.

[4] G. Marshall, J. L. Tapia, I. D. Apremont, C. Grandi, C. Barros, A. Alegria, J. Standen, R. Panizza, A. Bancalari, J. Lacarruba e J. Fabres, «Original Article A New Score for Predicting Neonatal Very Low Birth Weight Mortality Risk in the NEOCOSUR South American Network», *Journal of Perinatology*, pp. 577–582, 2005.

[5] M. M. Pollack, M. A. Koch, D. A. Bartel, I. Rapoport, R. Dhanireddy e A. A. E. El-mohandes, «A Comparison of Neonatal Mortality Risk Prediction Models in», *Pediatrics*, vol. 105, n.º 5, pp. 1051–1057, 2000.

[6] J. W. Lim, S.-h. Chung, D. R. Kang e C.-r. Kim, «Risk Factors for Cause-specific Mortality of Very-Low-Birth- Weight Infants in the Korean Neonatal Network», *Korean Medical Science*, n.º June 2014, S35–44, 2015.

[7] M. Cremer, S. Roll, C. Gräf, A. Weimann, C. Bührer e C. Dame, «Nucleated red blood cells as marker for an increased risk of unfavorable outcome and mortality in very low birth weight infants», *Early Human Development*, vol. 91, n.º 10, pp. 559–563, 2015.

[8] M. Aparecida, M. Gaiva, E. Fujimori, A. Paula e S. Sato, «Mortalidade neonatal em crianças com baixo peso ao nascer», *Rev Esc Enferm USP*, vol. 48, n.º 5, pp. 778–86, 2014.

[9] M. Cunha, A. Bettencourt, A. Almeida, G. Mimoso, P. Soares e T. Tomé, «O recém nascido de extremo baixo peso . Estado aos 2-3 anos . Resultados do Registo Nacional de Muito Baixo Peso de 2005 e 2006», *Acta Pediátrica Portuguesa*, vol. 44, n.º 1, pp. 1–8, 2013.

[10] J. Marôco, *Análise Estatística com o SPSS Statistics*. 2014, p. 990.

[11] A. Hall, C. Neves e A. Pereira, *Grande Maratona de Estatística no SPSS*. 2016, p. 312.

[12] M. A. Turkman e G. L. Silva, *Modelos Lineares Generalizados*. 2000, p. 151.

[13] A. DeMaris, *Regression With Social Data: Modeling Continuous and Limited Response Variables*. 1946.

[14] *XXS-Associação Portuguesa de Apoio ao Bebé Prematuro*. URL: `http://www.xxs-prematuros.com/`.

[15] L. Araujo e A. Reis, «Exame Físico Neonatal», em *Enfermagem na Prática Materno-Neonatal*, 2012, p. 312.

[16] A. O. Santos, «NIDCAP{®}: Uma filosofia de cuidados{...}», *Nascer e Crescer*, vol. XX, pp. 26–31, 2011.

[17] INE, *Estatísticas Demográficas 2015*. 2016, p. 178.

[18] Sociedade Portuguesa de Pediatria, «Prescrição Pré-natal de Corticoides para reduzir a Morbilidade e Mortalidade Neonatais», 2012.

[19] R. McCall, *Fundamental Statistics for the Behavioral Sciences*, 7th ed. Brooks/Cole Publishing Company, Pacific Grove, 1998.

[20] D. W. Hosmer e S. Lemeshow, *Applied Logistic Regression*, 2nd ed. John Wiley & Sons, New York, 2000.

[21] J. Barnes, *Statistical Analysis for Engineers and Scientist. A computer based aproach.* McGraw-Hill, New York, 1994.

[22] W. Vach, *Logistic Regression With Missing Values in the Covariates.* Springer-Verlag, 1994.

[23] C. S. d. P. Pereira, *O Aparecimento de Valores Omissos no Contexto da Regressão Logística*, 1996.

[24] R. Little e D. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. John Wiley & Sons, 1987.

[25] R. Veroneze, «Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla», Tese de Mestrado, Universidade Estadual de Campinas, 2011, p. 238.

[26] L. Beretta e A. Santaniello, «Nearest neighbor imputation algorithms: a critical evaluation», *BMC Medical Informatics and Decision Making*, vol. 16, n.º S3, p. 74, 2016.

[27] S. K. Sarkar, H. Midi e S. Rana, «Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study», *Journal of Applied Sciences*, vol. 1, pp. 26–35, 2011.

[28] M. Norušis, *SPSS 14.0 Advanced Statistical Procedures Companion.* Prentice Hall. New York, 2006.

[29] D. Pregibon, «Logistic Regression Diagnostics», *Annals of Statistics*, vol. 9, pp. 705–724, 1981.

[30] J. Hox, *Multilevel analysis: Techniques and applications*, 2nd ed. Erlbaum.Mahwah, NJ, 2010.

[31] S. P. Frutuoso, *Fórum Neonatal Português*, 2011. URL: `http://lusoneo.portugueseforum.net/`.

Anexo A

Critérios de Inclusão e Instruções de Preenchimento de 2010 da Base de Dados do Recém-Nascido de Muito Baixo Peso

CRITÉRIOS DE INCLUSÃO NA BASE DE DADOS

Qualquer recém-nascido (RN) vivo, nascido ou transferido para o hospital responsável pelo registo nas 1as 24 horas de vida, deve ser registado na base de dados desde que:

- Peso ao nascimento <1501 g, independentemente da idade gestacional (IG).

ou:

- IG menor que 32 semanas (31 + 6 inclusive), independentemente do peso.

ou:

- Gémeos de gémeos que cumpram os critérios acima.

A **responsabilidade do registo** é das unidades de cuidados intensivos/intermédios (UCIN) que tratam o RN. Se um RN nasceu em determinado hospital e foi transferido nas primeiras 24 horas de vida para outro hospital, a responsabilidade do registo é do hospital que recebe o RN e que o trata após as 24 horas de vida.

Quando um RN é transferido para outro hospital, após vários dias de tratamento numa unidade (ex: para crescimento, para opções terapêuticas específicas, para cirurgia, etc) antes de completar as 36 semanas de idade corrigida ou a observação por oftalmologia, o hospital que transfere e o hospital que recebe o RN têm que se articular de modo a não deixar por preencher estes dois itens.

INSTRUÇÕES DE PREENCHIMENTO DA BASE DE DADOS

Identificação

Nº Processo: Número do processo de internamento do RN.

Código: Campo de preenchimento automático, com código gerado pelo programa. Não é preciso registar nada.

A cada hospital é atribuído um código de identificação (CCC/RR/HH/YY/N):

CCC – Código do País - 351

RR – Código Regional

HH – Código do Hospital

YY – Últimos 2 dígitos do ano

NNN - Nº consecutivos de doentes admitidos

Data de nascimento: Data de nascimento do RN.

Hora de nascimento: Hora de nascimento do RN.

Nome da mãe: Nome completo da mãe do RN.

Idade da mãe: Idade em anos da mãe do RN.

Telefone da mãe: Telefone de contacto da mãe.

Código postal da mãe: Código postal da área de residência da mãe.

Nome da criança: Nome do RN.

Resumo do processo: Campo de preenchimento automático. Não é preciso registar nada.

Pré-admissão

Idade gestacional (IG):

- A melhor estimativa da IG, obtida no dia do nascimento, registada em semanas completas e dias.
- O nº de dias NÃO DEVE SER DEIXADO EM BRANCO (ex: se o RN nasceu com x semanas completas, colocar 0 no campo dos dias).

Peso ao nascer:

- Registar o primeiro peso obtido, em gramas. Se o RN faleceu na sala de partos e o único peso é o da autópsia, colocar o peso obtido na autópsia.

Comprimento:

- Registar o comprimento ao nascer, em centímetros, arredondado à décima.

Perímetro cefálico:

- Registar o perímetro cefálico ao nascer, em centímetros, arredondado à décima.

Morte na sala de partos:

- Registar "SIM", se o RN faleceu na sala de partos, antes da admissão na UCIN. Neste caso, completar a ficha com o preenchimento dos itens referentes aos procedimentos na sala de partos.
- Registar "NÃO", se o RN não faleceu na sala de partos. Neste caso, prosseguir com o preenchimento da folha de registo.

Local de nascimento:

- Registar "INBORN", se o RN nasceu no hospital responsável pelo registo.
- Registar "OUTBORN", se o RN nasceu fora do hospital de registo sendo para ele transferido.

Transferido de:

- Se o RN nasceu fora do hospital responsável pelo registo, escolher da listagem anexa o local de onde o RN veio transferido.

Transporte:

Transporte In-útero:

- Registar "SIM", se a mãe foi transferida de outra instituição hospitalar no período pré-parto, com o intuito do RN nascer no hospital responsável pelo registo.
- Registar "NÃO", se a mãe recorreu por moto próprio ao hospital responsável pelo registo, para o nascimento do RN.

Transporte Ex-útero:

- Registar "SIM", se o RN nasceu noutro local e foi transportado para o hospital responsável pelo registo nas primeiras 24 horas após o nascimento.
- Registar "NÃO", se o RN nasceu no hospital responsável pelo registo, não tendo sido sujeito a transporte nas primeiras 24 horas após o nascimento.

Data de admissão / hora de admissão na UCIN/UCERN:

- Data (dd/mm/aaaa) e hora (hh:mm) de admissão do RN na UCIN / UCERN. Os dias começam às 00:00h e terminam 23:59h.

Cuidados pré-natais:

- Registar "NÃO", se a mãe não recebeu cuidados obstétricos pré-natais antes da admissão para o parto.
- Registar "SIM", se a mãe recebeu cuidados obstétricos pré-natais.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Concepção assistida:

- Registar "NÃO", se a concepção não foi medicamente assistida.
- Registar "SIM", se a concepção foi medicamente assistida.

Corticóides pré-natais:

- Registar "NÃO", se não houve qualquer administração de corticóides antes do nascimento.

- Registar “PARCIAL”, se o nascimento ocorreu menos de 24 horas após a 1ª dose de corticóide, ou mais de uma semana após a última dose de corticóide.
- Registar “COMPLETO”, se o nascimento ocorreu mais de 24 horas e menos de uma semana, após pelo menos uma dose de corticóide.

Nº de ciclos:

- Registar o nº de ciclos de corticóide realizados.

Betametasona:

- Registar “NÃO”, se não foi administrada betametasona.
- Registar “SIM”, se foi administrada betametasona.

Dexametasona:

- Registar “NÃO”, se não foi administrada dexametasona.
- Registar “SIM”, se foi administrada dexametasona.

Patologias na Gravidez:

- Registar “NÃO”, se a gravidez decorreu sem patologia materna.
- Registar “SIM”, se foi detectada alguma patologia materna durante a gestação.

Descrição das patologias:

- Se foi detectada alguma patologia materna durante a gestação, descreve-la na caixa anexa
- Motivo do parto:
- Registar “ESPONTÂNEO”, se a mãe entrou espontaneamente em trabalho de parto.
- Registar “IVG”, se o parto ocorreu após tentativa frustre de interrupção voluntária da gravidez.
- Registar “PATOLOGIA MATERNA”, se o parto ocorreu após interrupção da gestação por patologia materna.
- Registar “PATOLOGIA FETAL”, se o parto ocorreu após interrupção da gestação por patologia fetal.

Sala de Partos

Tipo de Parto

- Registar "VAGINAL", para qualquer tipo de parto por via vaginal (espontâneo ou induzido).
- Registar "CESARIANA", para qualquer tipo de cesariana (electiva ou de emergência).

Sexo:

- Assinalar o sexo do RN: masculino, feminino ou indeterminado.

Gemelar:

- Registar "NÃO", se o RN resulta de gestação simples.
- Registar "SIM", se o RN resulta de qualquer tipo de gestação múltipla.

Total fetos:

- Na gestação múltipla, registar o nº total de fetos da gestação.

Nº de Ordem:

- Na gestação múltipla, registar o nº de ordem de nascimento do RN em questão (1, 2, ...).

Monocoriónico:

- Registar "NÃO", se a gestação gemelar é bi, tri, ... coriónica.
- Registar "SIM", se a gestação é monocoriónica.
- Registar "DÚVIDA", se não há dados disponíveis que permitam responder a esta questão.

Apgar:

- Registar o índice de Apgar ao 1º, 5º e 10º minutos.

Ressuscitação na sala de partos (corresponde à reanimação inicial, tenha sido ela realizada na sala de partos ou em qualquer outro local onde o RN tenha nascido - ex: bloco operatório, outro local hospitalar casa, ambulância, etc.):

Oxigénio:

- Registar "NÃO", se o RN não recebeu qualquer suplemento de oxigénio na sala de partos.
- Registar "SIM", se o RN recebeu qualquer suplemento de oxigénio na sala de partos.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Insuflador / Máscara:

- Registar "NÃO", se o RN não recebeu qualquer tipo de pressão positiva por máscara, *prongs* nasais ou via laríngea e insuflador manual, neopuff® ou similar na sala de partos

  "NÃO" igualmente, se máscara ou *prongs* nasais e insuflador manual, neopuff® ou similar foram usados para administrar apenas CPAP (pressão positiva contínua) não tendo sido administrada qualquer tipo de pressão positiva intermitente na sala de partos.
- Registar "SIM", se o RN recebeu qualquer tipo de pressão positiva por máscara ou *prongs* nasais e insuflador manual, neopuff® ou similar na sala de partos.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Entubação ET:

- Registar "NÃO", se o RN não foi entubado, ou se o tubo endotraqueal (TET) foi colocado apenas para aspiração não tendo sido submetido a ventilação assistida por TET na sala de partos.
- Registar "SIM", se o RN recebeu ventilação através de TET na sala de parto.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Compressão cardíaca:

- Registar "NÃO", se não foi efectuada massagem cardíaca externa na sala de partos.
- Registar "SIM", se foi efectuada massagem cardíaca externa na sala de partos.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Adrenalina:

- Registar "SIM", se foi ministrada adrenalina por qualquer via, na sala de partos.
- Registar "NÃO", se foi não foi ministrada adrenalina, na sala de partos
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Internamento

Suporte respiratório após a sala de partos:

Assinalar "SIM" ou "NÃO", consoante o RN tenha recebido ou não alguma das seguintes terapias respiratórias após a reanimação inicial na sala de partos:

- **Oxigénio**: suplemento de oxigénio, qualquer que tenha sido o modo de administração.
- **CPAP**: CPAP nasal.
- **VPPNI**: qualquer tipo de ventilação por pressão positiva não invasiva, sem entubação endotraqueal.
- **IPPV**: qualquer tipo de ventilação por pressão positiva via tubo endotraqueal.
- **VAF**: ventilação de alta frequência, via tubo endotraqueal.

Dias de ventilação:

- Assinalar o nº de dias em que o RN tenha recebido qualquer tipo de ventilação via TET. Assinalar como 1 desde que o RN tenha estado mais do que 1 hora ventilado via TET. Assinalar 0 se o RN não esteve ventilado via TET.

Surfactante:

Na sala de partos:

- Registar "NÃO", se o RN não recebeu surfactante exógeno na reanimação inicial (na sala de partos ou equivalente).
- Registar "SIM", se o RN recebeu surfactante exógeno durante a reanimação inicial (na sala de partos ou equivalente).
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Depois da sala de partos:

- Registar "NÃO", se o RN não recebeu surfactante exógeno para além de alguma dose dada na reanimação inicial (na sala de partos ou equivalente).

- Registar "SIM", se o RN recebeu uma ou mais doses de surfactante durante o internamento, para além de alguma dose administrada na reanimação inicial (na sala de partos ou equivalente).
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

Se, em algum dos 2 itens referentes ao surfactante, a resposta foi "SIM":

- Registar o **Nº de doses** de surfactante administradas.
- Registar a **Data da 1ª administração** (data da 1ª dose).
- Registar a **Hora da 1ª administração** (hora da 1ª dose).

CRIB:

- Abrir a janela e preencher os parâmetros assinalados. Não deixar nenhum parâmetro por assinalar. O CRIB é calculado automaticamente. Se não houver dados para preencher a totalidade dos campos, deixar em branco.

SNAPPE II

- Abrir a janela e preencher os parâmetros assinalados. Não deixar nenhum parâmetro por assinalar. O SNAPPE II é calculado automaticamente. Se não houver dados para preencher a totalidade dos campos, deixar em branco.

O2 no dia 28:

- Registar "NÃO", se o RN ainda estava hospitalizado, sem suplemento de O2 em dia 28 de vida.
- Registar "SIM", se o RN ainda estava hospitalizado e a receber qualquer suplemento de O2 em dia 28 de vida.
- Registar "NÃO APLICÁVEL", se o RN teve alta ou faleceu antes do Dia 28 de vida. Isto não é resposta do nosso programa. Devíamos substituir o desconhecido por não aplicável

O2 36 semanas:

- Registar "NÃO", se o RN ainda estava hospitalizado, sem suplemento de O2 às 36 semanas de idade corrigida.
- Registar "SIM", se o RN ainda estava hospitalizado e a receber qualquer suplemento de O2 às 36 semanas de idade corrigida.

- Registar "NÃO APLICÁVEL", se o RN teve alta ou faleceu antes de atingir as 36 semanas de idade corrigida, ou se o RN nasceu com idade gestacional próxima ou superior às 36 semanas.

Corticóides para DPC:

- Registar "NÃO", se não foram administrados corticóides após o nascimento para tratar ou prevenir doença pulmonar crónica / displasia broncopulmonar.
- Registar "SIM", se foram administrados corticóides após o nascimento para tratar ou prevenir doença pulmonar crónica / displasia broncopulmonar.

Diagnósticos:

SDR:

Definição de Síndrome de dificuldade respiratória (SDR):

- PaO2 <50 mmHg em ar ambiente, cianose central em ar ambiente ou necessidade de O2 suplementar para manter a PaO2 >50 mmHg.
- Radiografia do tórax compatível com SDR (volume pulmonar reduzido e padrão pulmonar reticulogranular, com ou sem broncograma aéreo).
- Registar "SIM", se o RN teve SDR definido como presença dos critérios A+B:
- Registar "NÃO", se o RN não cumpriu ambos os critérios "A" e "B".

Pneumotórax:

- Registar "NÃO", se o RN não teve ar extrapleural diagnosticado por radiografia ou drenagem pleural.
- Registar "SIM", se o RN teve ar extrapleural diagnosticado por radiografia ou drenagem pleural.

Para RN que tenham sido submetidos a cirurgia torácica e nos quais foi colocado um dreno torácico na altura da cirurgia, OU se foi detectado ar livre em radiografia torácica realizada imediatamente após a cirurgia sem necessidade de colocação de dreno torácico, assinalar "NÃO".

Para RN que tenham sido submetidos a cirurgia torácica e mais tarde desenvolveram ar extrapleural diagnosticado por radiografia do tórax ou drenagem pleural, assinalar "SIM".

PDA:

Definição de Persistência de ductos arteriosus (PDA) hemodinamicamente significativo (ecocardiografia): diâmetro transductal mínimo >1,5 mm; fluxo esquerdo – direito exclusivo contínuo; padrão de fluxo não restritivo através do canal arterial (velocidade sistólica máxima na extremidade pulmonar do canal arterial <2 m/s), sinais de hiperfluxo pulmonar e sobrecarga cardíaca esquerda (ø AE : Ao >1,5); sinais de hipoperfusão sistémica *(in "Consenso nacional de abordagem diagnóstica e terapêutica da persistência do canal arterial no RN pretermo" - Maio 2010).*

- Registar "NÃO", se o RN não teve PDA hemodinamicamente significativo.
- Registar "SIM", se o RN teve PDA hemodinamicamente significativo.
- Registar "DESCONHECIDO", se não há dados disponíveis que permitam responder a esta questão.

NEC:

Definição de Enterocolite Necrotizante (NEC):

A. Presença de um ou mais dos seguintes critérios clínicos: vómito ou aspirado gástrico biliar visível; distensão abdominal; fezes com sangue oculto ou visível; **E**

- Presença de um ou mais sinais radiológicos: pneumatosis intestinalis; ar hepatobiliar; pneumoperitoneu.
- Registar "NÃO", se o RN não cumpriu a definição de NEC.
- Registar "SIM", se o RN cumpriu a definição clínica e radiológica de NEC ou se teve diagnóstico de NEC no acto cirúrgico ou no exame pos-mortem.

Se o RN apresentou clínica e radiologia compatível com o diagnóstico de NEC, mas na cirurgia ou no exame pós-mortem se diagnosticou perfuração gastrointestinal focal, é este último diagnóstico que deve ser assinalado e não o de NEC.

Perfuração gastrointestinal (GI) focal:

- Registar "NÃO", se o RN não teve uma perfuração GI focal isolada independente de NEC, diagnosticada na cirurgia ou no exame pós-mortem.
- Registar "SIM", se o RN teve uma perfuração GI focal isolada independente de NEC, diagnosticada na cirurgia ou no exame pós-mortem.

Indometacina / Ibuprofeno (Profiláctico):

- Registar "NÃO", se não foi administrada indometacina ou ibuprofeno após o nascimento para profilaxia de PDA.
- Registar "SIM", se foi administrada indometacina ou ibuprofeno após o nascimento sem evidência de PDA.

Indometacina / Ibuprofeno (Terapêutico):

- Registar "NÃO", se não foi administrada indometacina ou ibuprofeno após o nascimento para tratamento de PDA.
- Registar "SIM", se foi administrada indometacina ou ibuprofeno após o nascimento para tratamento de PDA.

Cirurgia:

Laqueação PDA:

- Registar "NÃO", se não foi realizada laqueação cirúrgica do canal arterial.
- Registar "SIM", se foi realizada laqueação cirúrgica do canal arterial, na UCIN ou no bloco operatório.

Cirurgia NEC:

Realização de qualquer uma das seguintes intervenções para tratamento de enterocolite necrotizante (NEC), suspeita de NEC ou perfuração intestinal: laparotomia, ressecção intestinal ou colocação de dreno intraperitoneal.

- Registar "NÃO", se não foi realizada qualquer uma das intervenções mencionadas.
- Registar "SIM", se foi realizada uma ou mais das intervenções mencionadas.

Outra cirurgia major:

Realização de cirurgia major no bloco operatório ou na UCIN, para além de laqueação cirúrgica do canal arterial, cirurgia de NEC e cirurgia para tratamento de retinopatia da prematuridade (ROP). Os seguintes procedimentos não são considerados cirurgia major: piloromiotomia, herniorrafia uni ou bilateral, circuncisão e colocação de cateter central. Se forem realizadas laparotomias ou

ressecções intestinais múltiplas para NEC no período de uma semana, todos serão considerados "Cirurgia NEC" e apenas o item "Cirurgia NEC" deve ser assinalado.

- Registar "NÃO", se não foi realizada qualquer outra intervenção cirúrgica major, para além das mencionadas.
- Registar "SIM", se foi realizada qualquer outra intervenção cirúrgica major, para além das mencionadas. Neste caso, escolher da listagem anexa a cirurgia major realizada. Se escolheu "Outra cirurgia major" descreva-a no campo de "Observações gerais" no final do programa.

Imagiologia cerebral até ao dia 28:

- Registar "NÃO", se o RN não fez nenhum exame de imagem cerebral (Eco TF, RM ou TAC) até completar 28 dias de vida.
- Registar "SIM", se o RN fez pelo menos um exame de imagem cerebral (Eco TF, RM ou TAC) até

Nº Eco TF = 40 sem:

- Nº de ecografias TF que o RN realizou até às 40 semanas de IPC. Se não fez nenhuma, assinalar 0.
- "Sem informação", se não houver esta informação disponível.

Idade Eco TF/ RM + próxima das 40 sem.:

- Idade, em semanas completas, da Eco TF ou RMC que o RN tenha realizado mais próximo das 40 semanas. Se não fez nenhum dos exames mencionados ou se a idade da sua realização for desconhecida, assinalar "99".

Pior grau de HPIV (pior grau 0-3):

Se foi feito algum exame de imagem cerebral, registar o grau mais grave de hemorragia peri ou intraventricular (HIV) detectada, com base nos seguintes critérios:

- 0: sem evidência de HIV.
- 1: hemorragia da matriz germinal com ausência de hemorragia intraventricular (HIV) ou HIV < 10% da área ventricular.
- 2: HIV com 10-50% da área ventricular.

- 3: HIV >50% da área ventricular; habitualmente distende o ventrículo lateral.
- “Sem informação”, se não houver esta informação disponível.

EVHP:

Registar “NÃO”, se o RN não teve enfarte venoso hemorrágico periventricular (EVHP) associado à HIV.

Registar “SIM”, se o RN teve EVHP associado à HIV.

- Registar “SEM INFORMAÇÃO”, se não houver esta informação disponível.

EVHP: topografia:

- Escolher da listagem o(s) território(s) cerebral(ais) atingido(s) pelo EVHP
- Escolher “Sem informação”, se não houver esta informação disponível.

EVHP: extensão:

- “Unilateral sem desvio”, se o EVHP foi unilateral e não provocou desvio da linha média.
- “Unilateral com desvio”, se o EVHP foi unilateral e provocou desvio da linha média.
- “Bilateral sem desvio”, se o EVHP foi bilateral e não provocou desvio da linha média.
- “Bilateral com desvio”, se o EVHP foi bilateral e provocou desvio da linha média.
- “Sem informação”, se não houver esta informação disponível.

Dilatação ventricular pós-hemorrágica:

- Registar “NÃO”, se o RN não teve dilatação ventricular pós-hemorrágica.
- Registar “SIM”, se o RN teve dilatação ventricular pós-hemorrágica.
- Registar “SEM INFORMAÇÃO”, se não houver esta informação disponível.

LPV (pior grau):

Se foi feito algum exame de imagem cerebral, registar o grau mais grave de leucomalácia periventricular (LPV) detectada, com base nos seguintes critérios:

- 0: sem evidência de LPV.
- 1: hiperecogenicidade periventricular transitória persistindo =7 dias.
- 2: hiperecogenicidade periventricular transitória que evoluiu para pequenos quistos fronto-parietais localizados.

- 3: hiperecogenicidade periventricular que evoluiu para lesões quísticas periventriculares extensas.
- 4: hiperecogenicidade que atingiu a substância branca profunda, e que evoluiu para lesões quísticas extensas.
- "Sem informação", se não houver esta informação disponível.

LPV: data de diagnóstico:

- Data de diagnóstico da LPV.

LPV: idade de diagnóstico:

- Campo calculado e preenchido automaticamente, com a idade de diagnóstico, em dias, da LPV.

Outras alterações diagnosticadas / Observações:

- Campo de escrita livre para outras alterações detectadas ou observações pertinentes, referentes aos exames de imagem cerebral do RN.

Sépsis e/ou meningite precoce (£D3):

Definição: se o RN teve, nas primeiras 72 horas de vida, clínica compatível com sépsis e/ou meningite, tratamento ou intenção de tratamento antibiótico pelo menos durante 5 dias e agente bacteriano responsável isolado em hemocultura ou cultura de líquor. Se não foi isolado agente nas culturas mencionadas, mas o RN teve clínica e parâmetros analíticos compatíveis com sépsis e/ou meningite e tratamento ou intenção de tratamento antibiótico pelo menos durante 5 dias, considerar sépsis e/ou meningite sem agente identificado.

- Registar "NÃO", se o RN não teve diagnóstico compatível com sépsis e/ou meningite precoce.
- Registar "SIM", se o RN teve diagnóstico compatível com sépsis e/ou meningite precoce.

Agente (cod.):

- Escolher da listagem o agente responsável pela sépsis e/ou meningite precoce. Se o RN teve este diagnóstico mas não foi isolado agente nas culturas mencionadas, escolher "Sem agente identificado".

Sépsis e/ou meningite tardia (>D3):

Definição: se o RN teve, após as 72 horas de vida, clínica compatível com sépsis e/ou meningite, tratamento ou intenção de tratamento antibiótico pelo menos durante 5 dias e agente bacteriano responsável isolado em hemocultura ou cultura de líquor. Se não foi isolado agente nas culturas mencionadas, mas o RN teve clínica e parâmetros analíticos compatíveis com sépsis e/ou meningite e tratamento ou intenção de tratamento antibiótico pelo menos durante 5 dias, considerar sépsis e/ou meningite sem agente identificado.

- Registar "NÃO", se o RN não teve diagnóstico compatível com sépsis e/ou meningite tardia.
- Registar "SIM", se o RN teve diagnóstico compatível com sépsis e/ou meningite tardia.

Nº Episódio / Agente:

- Para cada episódio de sépsis e/ou meningite tardia, escolher da listagem o nº sequencial do episódio infeccioso e o respectivo agente responsável. Se o RN teve este diagnóstico mas não foi isolado agente nas culturas mencionadas, escolher "Sem agente identificado".

Exame oftalmológico:

- Registar "NÃO", se o RN não foi submetido a exame oftalmológico com observação da retina.
- Registar "SIM", se o RN foi submetido a exame oftalmológico com observação da retina.

ROP – Pior grau (0-5):

Se a resposta ao Exame Oftalmológico foi "SIM", especificar o pior grau de retinopatia da prematuridade (ROP), de acordo com a seguinte classificação:

- 0: Sem evidência de lesões compatíveis com ROP.
- 1: linha de demarcação que separa a retina posterior vascularizada da anterior avascular.
- 2: prega ou linha de demarcação espessa.
- 3: prega com proliferação brovascular extra-retiniana.
- 4: descolamento parcial da retina.
- 5: descolamento total da retina.

Doença " Plus":

- Registar "SIM", se foi diagnosticada ROP grau 2 ou 3, associadas a sinais de incompetência vascular (dilatação progressiva e tortuosidade vascular).
- Registar "NÃO", em todos os outros casos.

Cirurgia ROP:

- Registar "NÃO", se o RN não foi submetido a crio-cirurgia ou a tratamento laser para ROP.
- Registar "SIM", se o RN foi submetido a crio-cirurgia ou a tratamento laser para ROP.

Malformação congénita major:

- Registar "NÃO", se não foi diagnosticada ao RN qualquer malformação congénita major.
- Registar "SIM", se foi diagnosticada ao RN uma ou mais malformações congénitas major. Se sim, registar a(s) malformação(ões) detectada(s) na caixa anexa, escolhendo-as da listagem fornecida. Se escolheu "Outra Malformação Congénita Letal ou Ameaçadora de Vida" descreva-a no campo de "Observações gerais" no final do programa.

Destino

Óbito:

- Registar "NÃO, se o RN não faleceu durante o internamento neonatal".
- Registar "SIM", se o RN faleceu durante o internamento neonatal".

Data:

- Data de óbito

Hora:

- Hora de óbito.

Causa da morte:

- Escolher a causa de morte principal da listagem anexa.
- Se for escolhida a opção "Outra causa de morte", descreva-a no campo de "Observações gerais" no final do programa.

Autópsia:

- Registar "NÃO", se não foi realizado exame necrópsico.
- Registar "SIM", se foi realizado exame necrópsico.

Abstenção de cuidados terapêuticos:

- Registar "NÃO", se não foi tomada qualquer decisão de suspensão de cuidados em curso ou de iniciar novas terapêuticas curativas.
- Registar "SIM", se foi tomada decisão de suspensão de cuidados em curso ou de não iniciar novas terapêuticas curativas, em RN cuja hipótese de sobrevivência tenha sido considerada mínima.

Transferido:

- Registar "NÃO", se o RN não foi transferido para outro(s) hospital(ais) antes de completar 1 ano de idade e antes de alguma vez ter tido alta para o domicílio.
- Registar "SIM", se o RN foi transferido para outro(s) hospital(ais) antes de completar 1 ano de idade e antes de alguma vez ter tido alta para o domicílio.
- Caso o RN tenha tido uma ou mais transferências, registar na caixa anexa a(s) data(s) da(s) transferência(s), o(s) Hospital (ais) para onde foi transferido, e o(s) motivo(s) da transferência(s) escolhidos da listagem anexa. Se tiver sido escolhida a opção "Outra causa de transferência", descreva-a no campo de "Observações gerais" no final do programa.

Dados na altura da 1ª transferência:

Alimentação entérica:

- Registar "NENHUMA", se o RN não estava a receber nenhum tipo de alimentação entérica, na altura da 1ª transferência.

- Registar "LEITE MATERNO", se o RN estava a receber como alimentação entérica, unicamente leite materno não fortificado, na altura da 1ª transferência.
- "LEITE MATERNO COM FORTIFICANTE", se o RN estava a receber como alimentação entérica, unicamente leite materno fortificado, na altura da 1ª transferência.
- "LEITE DE FÓRMULA", se o RN estava a receber como alimentação entérica, unicamente leite de fórmula, na altura da 1ª transferência.
- "OS DOIS", se o RN estava a receber como alimentação entérica leite materno (fortificado ou não) associado a leite de fórmula, na altura da 1ª transferência.

Peso:

- Registar o peso, em gramas, do dia da transferência ou do dia anterior.

Comprimento:

- Registar o comprimento, em centímetros, arredondado à décima, do dia da transferência ou do dia anterior.

Perímetro cefálico:

- Registar o perímetro cefálico, em centímetros, arredondado à décima, do dia da transferência ou do dia anterior.

Dependência de O2:

- Registar "NÃO", se à data da 1ª transferência o RN não necessitava de suplemento de O2.
- Registar "SIM", se à data da 1ª transferência o RN necessitava de suplemento de O2.

Monitor de apneia:

- Registar "NÃO", se à data da 1ª transferência o RN não necessitava de monitor de apneia.
- Registar "SIM", se à data da 1ª transferência o RN necessitava de monitor de apneia.

Domicílio:

- Registar "NÃO", se o RN nunca teve alta para o domicílio até completar 1 ano de idade.

- Registar “SIM”, se o RN teve alta para o domicílio até completar 1 ano de idade.

Data:

- Se o RN teve alta para o domicílio até completar 1 ano de idade, preencher na caixa a data de alta para o domicílio.

Idade:

- Idade, em dias, à data de alta para o domicílio (campo de cálculo automático).

Internamento ao ano de idade:

- Registar “NÃO”, se o RN teve alta para o domicílio antes de cumprir 365 dias de idade.
- Registar “SIM”, se a criança esteve consecutivamente internada, permanecendo ainda internada ao cumprir 365 dias de idade.

Estado final:

Dados das 24 horas que precederam a alta para o domicílio, o óbito ou ao cumprir 1 ano de idade se ainda internado.

Alimentação entérica:

- Registar “NENHUMA”, se o RN não estava a receber nenhum tipo de alimentação entérica.
- Registar “LEITE MATERNO”, se o RN estava a receber como alimentação entérica, unicamente leite materno não fortificado.
- Registar “LEITE MATERNO COM FORTIFICANTE”, se o RN estava a receber como alimentação entérica, unicamente leite materno fortificado.
- Registar “LEITE DE FÓRMULA”, se o RN estava a receber como alimentação entérica, unicamente leite de fórmula.
- Registar “OS DOIS”, se o RN estava a receber como alimentação entérica leite materno (fortificado ou não) associado a leite de fórmula.

Complicações:

Ausentes:

- Registar "NÃO", se o RN apresentava algum tipo de complicação ou sequela à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.
- Registar "SIM", se o RN não apresentava nenhum tipo de complicação ou sequela à data da alta para o domicílio.

Respiratórias:

- Registar "NÃO", se o RN não apresentava nenhum tipo de complicação ou sequela respiratória à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.
- Registar "SIM", se o RN apresentava alguma complicação ou sequela respiratória à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Dependência de O2:

- Registar "NÃO", se o RN não necessitava de suplemento de O2 à data da alta para o domicílio, do óbito ou se ainda internado ao cumprir 1ano de idade.
- Registar "SIM", se o RN necessitava de suplemento de O2 à data da alta para o domicílio, do óbito ou se ainda internado ao cumprir 1ano de idade.

Monitor de apneia:

- Registar "NÃO", se o RN não necessitava de monitor de apneia à data da alta para o domicílio, do óbito ou se ainda internado ao cumprir 1ano de idade.
- Registar "NÃO", se o RN necessitava de monitor de apneia à data da alta para o domicílio, do óbito ou se ainda internado ao cumprir 1ano de idade.

Digestivas:

- Registar "NÃO", se o RN não apresentava nenhum tipo de complicação ou sequela digestiva à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.
- Registar "SIM", se o RN apresentava alguma complicação ou sequela digestiva à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Neurológicas:

- Registar "NÃO", se o RN não apresentava nenhum tipo de complicação ou sequela neurológica à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.
- Registar "SIM", se o RN apresentava alguma complicação ou sequela neurológica à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Hidrocefalia:

Campo disponibilizado apenas se foi detectada ao RN dilatação ventricular pós-hemorrágica.

- Registar "NÃO", se o RN não teve o diagnóstico de hidrocefalia.
- Registar "SEM INFORMAÇÃO", se não houver esta informação disponível.
- Das restantes opções (Estável sem drenagem / Drenagem transitória / Drenagem definitiva / Drenagem transitória + definitiva), registar a adequada à situação final do RN.

Outras:

- Registar "NÃO", se o RN não apresentava nenhum outro tipo de complicação ou sequela à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.
- Registar "SIM", se o RN apresentava algum outro tipo de complicação ou sequela à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado. Neste caso, assinalar as complicações ou sequelas detectadas na caixa anexa.

Observações:

Caixa de escrita livre para referir dados considerados de interesse, ou que ajudem a esclarecer algum dado registado, referente ao estado final do RN em causa

Antropometria final:

Peso:

- Registar o peso, em gramas, à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Comprimento:

- Registar o comprimento, em centímetros, arredondado à décima, à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Perímetro cefálico:

- Registar o perímetro cefálico, em centímetros, arredondado à décima, à data da alta para o domicílio, do óbito ou ao cumprir 1 ano de idade se ainda internado.

Hospital de seguimento da criança:

- Registar o Hospital responsável pelo seguimento da criança.

Nº Processo:

- Registar o nº do processo do Hospital de Seguimento da criança

Estado da ficha:

- Registar "ABERTA", se o médico registador considerar a ficha inacabada. No caso de um RN ter sido transferido para outra instituição hospitalar, a ficha deve manter-se aberta até que os dados referentes à alta para o domicílio, internamento ao ano de idade e estado final tenham sido concluídos pelo registador do último hospital a receber a criança.
- Registar "FECHADA", se o médico registador considerar a ficha terminada, mesmo que ainda tenha campos por preencher. Neste caso o registador assume que não terá mais oportunidade de obter os dados em falta, ficando estes omissos no registo final.

Observações gerais:

- Campo de escrita livre para observações pertinentes que não tenham sido escritas noutros campos do registo.

Anexo B

Recodificação da Base de Dados Ficha associada aos Dados do Recém-Nascido de Muito Baixo Peso

Ficha

Coluna	Domínio
FichaId	<número> -999=(Não Aplicável)
UnidadeSaude	<Nome da Unidade de Saúde>
CentroHospitalar	<Texto livre>
RegiaoSaude	<Texto livre>
Processo	<número> -999=(Não Aplicável)
Codigo	<número> -999=(Não Aplicável)
DataNascimento	<Data e hora> (1973-01-01 = Não Aplicável)
Nome	<Texto livre>
MaeNome	<Texto livre>
MaeIdade	<número> -999=(Não Aplicável)
MaeTelefone	<número> -999=(Não Aplicável)
MaeCodigoPostal	<Texto livre>
MaeMorada	<Texto livre>
IdadeGestacional	<número>
NascimentoPeso	<número>
NascimentoComprimento	<número> -999=(Não Aplicável)
NascimentoPerimetroCefalico	<número> -999=(Não Aplicável)
NascimentoOutborn	1 = Outborn 2 = Inborn
NascimentoTipoLocal	1 = Hospital de Apoio Perinatal Diferenciado, 2 = Hospital de Apoio Perinatal 3 = Instituição de Saúde sem Apoio Perinatal 4 = Local extra Hospitalar 5 = Hospital Privado
NascimentoUnidadeSaude	<Texto livre>
Transporte	1 = In-Utero 2 = Ex-Utero 3 = Admissão Materna Directa
CuidadosPrenatais	1 = Sim 2 = Não 999 = (Não Aplicável)

ConcepcaoAssistida	1 = Sim 2 = Não 999 = (Não Aplicável)
CorticoidesPrenatais	1 = Não 2 = Parcial 3 = Completo 4 = Desconhecido 999=(Não Aplicável)
CorticoidesPrenataisCiclos	<número> -999=(Não Aplicável)
CorticoidesPrenataisUsado	<número> -999=(Não Aplicável)
PatologiasNaGravidez	1 = Sim 2 = Não 999 = (Não Aplicável)
TipoDeParto	1=Vaginal 2=Cesariana
MotivoDoParto	1=Espontâneo 2=Patologia materna 3=Patologia Fetal 4=IVG
Sexo	1=Masculino 2=Feminino 3=Indeterminado
Gemelar	1 = Sim 2 = Não 999 = (Não Aplicável)
GemelarOrdem	<número> -999=(Não Aplicável)
GemelarTotal	<número> -999=(Não Aplicável)
GemelarCorionicidade	1=Monocoriónico 2=Bi/Tricoriónico 9=Desconhecido 999=(Não Aplicável)
Apgar1	<número> -999=(Não Aplicável)
Apgar5	<número> -999=(Não Aplicável)
Apgar10	<número> -999=(Não Aplicável)
RessuscitacaoOxigenio	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)

RessuscitacaoInsuflador	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
RessuscitacaoEntubacaoEt	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
RessuscitacaoCompressaoCardiaca	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
RessuscitacaoAdrenalina	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
MorteNaSalaDePartos	1 = Sim 2 = Não 999 = (Não Aplicável)
DataAdmissaoUci	<Data e hora> (1973-01-01 = Não Aplicável)
IdadeNaAdmissaoEmMinutos	<número> -999=(Não Aplicável)
IdadeNaAdmissaoEmHoras	<número> -999=(Não Aplicável)
SROxigenio	1 = Sim 2 = Não 999 = (Não Aplicável)
SRCpap	1 = Sim 2 = Não 999 = (Não Aplicável)
SRVppni	1 = Sim 2 = Não 999 = (Não Aplicável)
SRVentilacaoOxidoNitrico	1 = Sim 2 = Não 999 = (Não Aplicável)
SRIppv	1 = Sim 2 = Não 999 = (Não Aplicável)
SRVaf	1 = Sim 2 = Não 999 = (Não Aplicável)
SRVafni	1 = Sim 2 = Não 999 = (Não Aplicável)

SRDiasVentilacao	<número> -999=(Não Aplicável)
MalformacaoCongenitaMajor	1 = Sim 2 = Não 999 = (Não Aplicável)
SepsisMeningiteTardia	1 = Sim 2 = Não 999 = (Não Aplicável)
SurfactanteInicial	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
SurfactantePosterior	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
SurfactantePosteriorDoses	<número> -999=(Não Aplicável)
SurfactantePosteriorData	<Data e hora> (1973-01-01 = Não Aplicável)
SurfactanteCrib	<número> 999 = Desconhecido -999=(Não Aplicável)
SurfactanteSnappe2	<número> 999 = Desconhecido -999=(Não Aplicável)
OxigenioDia28	1 = Sim 2 = Não 999 = Não Aplicável
OxigenioSemana36	1 = Sim 2 = Não 999 = Não Aplicável
CorticoidesDPC	1 = Sim 2 = Não 999 = (Não Aplicável)
DiagSdr	1 = Sim 2 = Não 999 = (Não Aplicável)
DiagPneumotorax	1 = Sim 2 = Não 999 = (Não Aplicável)
DiagPda	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)

	1 = Sim 2 = Não
DiagNec	999 = (Não Aplicável)
	1 = Sim 2 = Não
DiagPerfuracaoGi	999 = (Não Aplicável)
	1 = Sim 2 = Não
PdaProfilatico	999 = (Não Aplicável)
	1 = Sim 2 = Não
PdaTerapeutico	999 = (Não Aplicável)
	1 = Sim 2 = Não
CirurgiaPda	999 = (Não Aplicável)
	1 = Sim 2 = Não
CirurgiaNec	999 = (Não Aplicável)
	1 = Sim 2 = Não
CirurgiaMajorOutra	999 = (Não Aplicável)
	1 = Sim, 2 = Não, 9=Desconhecido
ImagiologiaCerebralDia28	999 = (Não Aplicável)
	0=0 1=1 2=2 3=3 10=4 ou mais
EcografiaTf	999=(Não Aplicável)
	<número>
EcografiaTfIdadeUltimaSemana	-999=(Não Aplicável)
	0=0 1=1 2=2 3=3 9=Desconhecido
Hpiv	999=(Não Aplicável)
	1 = Sim, 2 = Não, 9=Desconhecido
Evhp	999 = (Não Aplicável)

EvhpExtensao	1=Unilateral sem desvio 2=Unilateral com desvio 3=Bilateral sem desvio 4=Bilateral com desvio 9=Desconhecido 999=(Não Aplicável)
DilatacaoVentricularPh	1 = Sim, 2 = Não, 9=Desconhecido 999 = (Não Aplicável)
LpvGrau	0=0 1=1 2=2 3=3 4=4 9=Sem Informação 999=(Não Aplicável)
LpvDataDiagnostico	<Data sem hora> 1973-01-01 = (Não Aplicável)
LpvTexto	<Texto livre>
IdadeLpvDataDiagnosticoEmDias	<número> -999=(Não Aplicável)
SepsisMeningitePrecoce	1 = Sim 2 = Não 999 = (Não Aplicável)
SepsisMeningitePrecoceAgente	<Nome do Agente>
ExameOftalmologico	1 = Sim 2 = Não 999 = (Não Aplicável)
ExameOftalmologicoRopGrau	0=0 1=1 2=2 3=3 4=4 5=5 999=(Não Aplicável)
ExameOftalmologicoRopCirurgia	1 = Sim 2 = Não 999 = (Não Aplicável)
ExameOftalmologicoPlus	1 = Sim 2 = Não 999 = (Não Aplicável)
ObitoData	<Data sem hora> 1973-01-01 = (Não Aplicável)
ObitoAutopsia	1 = Sim 2 = Não 999 = (Não Aplicável)

ObitoCausa	1=Causa neurológica 2=Insuficiência respiratória 3=Malformação congénita 4=Sépsis/Outra Infeção 5=Outra causa de morte 6=Causa desconhecida
ObitoAbstencaoCuidadosTerapeuticos	1 = Sim 2 = Não 999 = (Não Aplicável)
IdadeObitoEmDias	<número> -999=(Não Aplicável)
IdadeObitoEmHoras	<número> -999=(Não Aplicável)
TransferenciaAlimentacaoEnterica	1=Nenhuma 2=Leite materno 3=Leite materno com fortificante 4=Leite de fórmula 5=Os dois 999=(Não Aplicável)
TransferenciaPeso	<número> -999=(Não Aplicável)
TransferenciaComprimento	<número> -999=(Não Aplicável)
TransferenciaPerimetroCefalico	<número> -999=(Não Aplicável)
TransferenciaDependenciaOxigenio	1 = Sim 2 = Não 999 = (Não Aplicável)
TransferenciaMonitorApneia	1 = Sim 2 = Não 999 = (Não Aplicável)
TransferenciaDestino	1=Domicílio 3=Óbito 4=Internado com 1 ano de idade 999=(Não Aplicável)
InternamentoAlimentacaoEnterica	1=Nenhuma 2=Leite materno 3=Leite materno com fortificante 4=Leite de fórmula 5=Os dois 999=(Não Aplicável)
InternamentoPeso	<número> -999=(Não Aplicável)
InternamentoComprimento	<número> -999=(Não Aplicável)
InternamentoPerimetroCefalico	<número> -999=(Não Aplicável)

InternamentoDependenciaOxigenio	1 = Sim 2 = Não 999 = (Não Aplicável)
InternamentoMonitorApneia	1 = Sim 2 = Não 999 = (Não Aplicável)
InternamentoDestino	1=Domicílio 2=Transferência 3=Óbito 4=Internado com 1 ano de idade 999=(Não Aplicável)
DataDestinoFinal	<Data sem hora> 1973-01-01 = (Não Aplicável)
IdadeDataDestinoEmDias	<número> -999=(Não Aplicável)
DescricaoComplicacao	<Texto livre>
Observacoes	<Texto livre>
HospitalSeguimento	<Nome da Unidade de Saúde>
HospitalSeguimentoProcesso	<número> -999=(Não Aplicável)
Estado	1=Aberta 2=Fechada